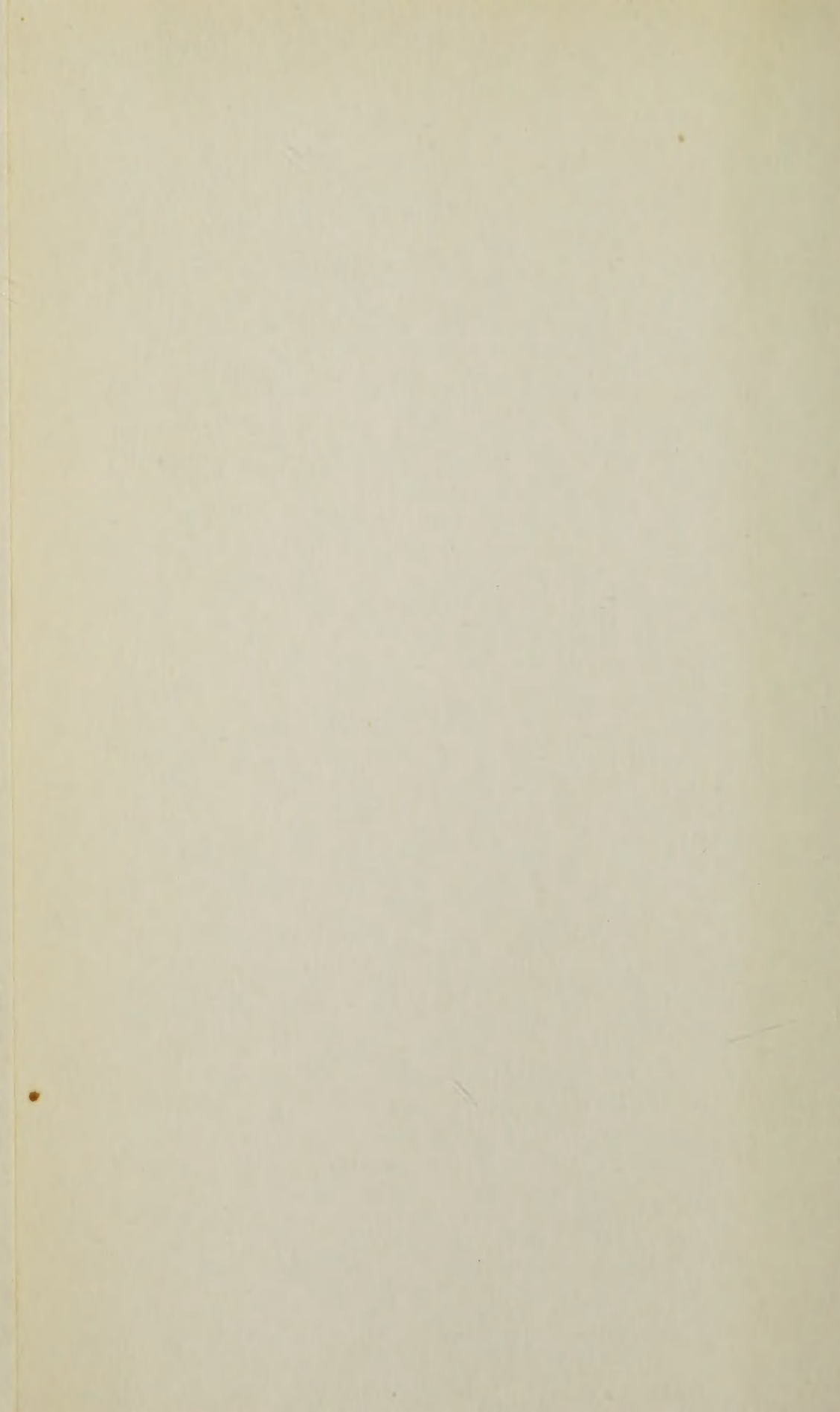


Digitized by the Internet Archive
in 2020 with funding from
Kahle/Austin Foundation

150.5

4812

v. 30



Journal of Applied Psychology

Edited by

Donald G. Paterson

University of Minnesota

Consulting Editors

PAUL S. ACHILLES, *Psychological Corporation*; WALTER V. BINGHAM, *A.G.O., War Department*; HAROLD E. BURTT, *Ohio State University*; ARTHUR I. GATES, *T. C. Columbia University*; JOHN G. JENKINS, *University of Maryland*; IRVING LORGE, *T. C. Columbia University*; QUINN MCNEMAR, *Stanford University*; WILLARD C. OLSON, *University of Michigan*; JAMES P. PORTER, *Swarthmore, Pennsylvania*; EDWARD K. STRONG, JR., *Stanford University*; MORRIS S. VITELES, *University of Pennsylvania*; JOSEPH ZUBIN, *N. Y. Psychiatric Institute*.

Volume 30, 1946

Published Bi-monthly by The American Psychological Association, Inc.
Prince and Lemon Sts., Lancaster, Pa., and 1515 Massachusetts Ave., NW, Washington 5, D. C.

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the Act of March 3, 1879.
Copyright, 1946, by The American Psychological Association, Inc.

Contents of Volume 30

Articles

Altus, W. D. The Comparative Validities of Two Tests of General Aptitude in an Army Special Training Center.....	42
Altus, W. D. and Mahler, C. A. The Significance of Verbal Aptitude in the Type of Occupation Pursued by Illiterates.....	155
Barrett, D. M. Prediction of Achievement in Typewriting and Stenography in a Liberal Arts College.....	624
Bauman, M. K. Studies in the Application of Motor Skills Techniques to the Vocational Adjustment of the Blind.....	144
Baxter, B. and Potechin, E. A Simplified Form for Reporting Test Results.....	32
Bolanovich, D. J. Statistical Analysis of an Industrial Rating Chart	23
Brozek, J., Guetzkow, H., Mickelsen, O., and Keys, A. Motor Performance of Normal Young Men Maintained on Restricted Intakes of Vitamin B Complex.....	359
Buck, J. N. Time Appreciation Test.....	388
Davis, J. Correlation Between Scores on Ortho-Rater Tests and Clinical Tests.....	596
Dimmick, F. L. A Color Aptitude Test, 1940 Experimental Edition	10
Drake, L. E. A Social I. E. Scale for the Minnesota Multiphasic Personality Inventory.....	51
Edwards, A. L. and Kenney, K. C. A Comparison of the Thurstone and Likert Techniques of Attitude Scale Construction.....	72
File, Q. W. and Remmers, H. H. Studies in Supervisory Evaluation	421
Fischer, R. P. Signed Versus Unsigned Personal Questionnaires..	220
Fisher, M. B. and Birren, J. E. Standardization of a Test of Hand Strength.....	380
Foster, H. A Comparative Study of Three Tests for Color Vision..	135
Giese, W. J. The Interrelationship of Visual Acuity at Different Distances.....	91
Goodman, C. H. The MacQuarrie Test for Mechanical Ability: I. Selecting Radio Assembly Operators.....	586
Horrocks, J. E. The Relationship Between Knowledge of Human Development and Ability to Use Such Knowledge.....	501
Humm, D. G. Test Validation on Remote Criteria.....	333
Keller, F. S., Christo, I. J., and Schoenfeld, W. N. Studies in International Morse Code: V. The Effect of the "Phonetic Equivalent"	265

Kellogg, W. N. The Learning Curve for Flying an Airplane	435
Kerr, W. A. Attitudes Toward Scheduling of Industrial Music	575
Kilby, R. W. Relation of Iowa Silent Reading Test Scores to Measures of Aptitude and Achievement	399
Lawshe, C. H., Jr. and Alessi, L. Studies in Job Evaluation: IV. Analysis of Another Point Rating Scale for Hourly-Paid Jobs and the Adequacy of an Abbreviated Scale	310
Lawshe, C. H., Jr. and Maleski, A. A. Studies in Job Evaluation: III. An Analysis of Point Ratings for Salary Paid Jobs in an Industrial Plant	117
Lawshe, C. H., Jr. Semanek, I. A. and Tiffin, J. The Purdue Mechanical Adaptability Test	442
Lawshe, C. H., Jr. and Wilson, R. F. Studies in Job Evaluation: V. An Analysis of the Factor Comparison System as it Functions in a Paper Mill	426
Lehman, H. C. Age of Starting to Contribute versus Total Creative Output	460
Lindner, R. M. and Gurvitz, M. Restandardization of the Revised Beta Examination to Yield the Wechsler Type of IQ	649
Link, H. C. Psychological Corporation's Index of Public Opinion	297
Link, H. C. The Psychological Corporation's Index of Public Opinion	1
Lough, O. M. Teachers College Students and the Minnesota Multiphasic Personality Inventory	241
Malamud, R. F. Validity of the Hunt-Minnesota Test for Organic Brain Damage	271
Manson, M. P. and Grayson, H. M. Keysort Method of Scoring the Minnesota Multiphasic Personality Inventory	509
McClure, J. The Development and Standardization of a New Type Test of Peripheral Vision	340
McMurry, R. N. Management's Reactions to Employee Opinion Polls	212
McNamara, W. J. and Weitzman, E. The Economy of Item Analysis with the IBM Graphic Item Counter	84
Meehl, P. E. Profile Analysis of the Minnesota Multiphasic Personality Inventory in Differential Diagnosis	517
Meehl, P. E. and Hathaway, S. R. The K Factor as a Suppressor Variable in the Minnesota Multiphasic Personality Inventory	525
Meehl, P. E. and Jeffery, M. The Hunt-Minnesota Test for Organic Brain Damage in Cases of Functional Depression	276
Mellenbruch, P. L. A Preliminary Report on the Miami-Oxford Curve-Block Series	129

Paterson, D. G. and Tinker, M. A. Readability of Newspaper Headlines Printed in Capitals and in Lower Case.....	161
Paterson, D. G. and Tinker, M. A. The Relative Readability of Newsprint and Book Print.....	454
Peixotto, H. E. The Relationship of College Board Examination Scores and Reading Scores for College Freshmen.....	406
Pressey, S. L. Age of College Graduation and Success in Adult Life	226
Rabin, A. I., Davis, J. C. and Sanderson, M. H. Item Difficulty of Some Wechsler-Bellevue Subtests.....	493
Rogers, R. C. Analysis of Two Point-Rating Job Evaluation Plans	579
Rohde, A. R. Explorations in Personality by the Sentence Completion Method.....	169
Rothe, H. F. Output Rates Among Butter Wrappers: I. Work Curves and Their Stability.....	199
Rothe, H. F. Output Rates Among Butter Wrappers: II. Frequency Distributions and an Hypothesis Regarding the "Restriction of Output".....	320
Sartain, A. Q. Predicting Success in ■ School of Nursing.....	234
Sartain, A. Q. Relation Between Scores on Certain Standard Tests and Supervisory Success in Aircraft Factory.....	328
Smith, G. M. The Effect of Prolonged Mild Anoxia on Speech Intelligibility.....	255
Smith, G. M., and Seitz, C. P. Speech Intelligibility Under Various Degrees of Anoxia.....	182
Strong, E. K., Jr. Interests of Senior and Junior Public Administrators.....	55
Stuit, D. B. and Wilson, J. T. The Effect of an Increasingly Well Defined Criterion on the Prediction of Success at Naval Training School (Tactical Radar).....	614
Tinker, M. A. and Paterson, D. G. Readability of Mixed Type Forms.....	631
Tuckman, J. A Comparison of the Reliability and Performance for the Minnesota Rate of Manipulation Test for Subjects Tested Individually and in Groups of Two.....	37
Tuckman, J. The Relationship Between Subjective Estimates of Personal Adjustment and Ratings on the Bell Adjustment Inventory.....	488
Verniaud, W. M. Occupational Differences in the Minnesota Multiphasic Personality Inventory.....	604
Weitz, R. D. The Occupational Adjustment Characteristics of ■ Group of Sexually Promiscuous and Venereally Infected Females	248
Welch, L. Recombination of Ideas in Creative Thinking.....	638

Winfield, M. C. The Use of the Harrower-Erickson Multiple Choice Rorschach Test with a Selected Group of Women in Military Service.....	481
Wirt, S. E. Statistical Laboratory for Vision Tests at Purdue University.....	354
Wright, M. E. Use of the Shipley-Hartford Test in Evaluating Intellectual Functioning of Neuropsychiatric Patients.....	45

Book Reviews

Beaumont's The Psychology of Personnel: Charles C. Gibbons.....	568
Bills' The Psychology of Efficiency. A Discussion of the Hygiene of Mental Work: Forrest A. Kingsbury.....	114
Boring's Psychology for the Armed Services: Albert T. Poffenberger	569
Brandt's The Psychology of Seeing: Miles A. Tinker.....	417
Cantor's Employee Counseling: Charles C. Gibbons.....	109
Chamberlin, Chamberlin, Drought and Scott's Did They Succeed in College?: E. G. Williamson.....	291
De Silva's Why Do We Have Automobile Accidents?: A. R. Lauer.	194
Dicks' Pastoral Work and Personal Counseling: J. Gustav White..	111
Gann's Reading Difficulty and Personality: Miles A. Tinker.....	571
Gardner's Human Relations in Industry: Douglas McGregor.....	413
Hahn and Brayfield's Occupational Laboratory Manual and Science Research Associates' Job Exploration Workbook: Donald E. Super.....	107
Hayes' Vocational Aptitude Tests for the Blind: S. G. DiMichael..	663
Hudson and Fish's Occupational Therapy in the Treatment of the Tuberculous Patient: Gwendolen G. Schneider.....	112
Inbau's Lie Detection and Criminal Interrogation: Howard F. Hunt	193
Lazarsfeld, Berelson, and Gaudet's The People's Choice. How the Voter Makes Up His Mind in a Presidential Campaign: Alfred C. Welch.....	289
Lowenfeld's Braille and Talking Book Reading: A Comparative Study: Samuel P. Hayes.....	665
Munroe's Prediction of the Adjustment and Academic Performance of College Students by a Modification of the Rorschach Method: Paul E. Meehl.....	660
Nesbitt's The Road to Avalon and Barton's And Now to Live Again: Horace B. English.....	570
NORC Interview Department's Interviewing for NORC: Philip H. Kriedt.....	659
Peatman and Hallonquist's The Patterning of Listener Attitudes toward Radio Broadcasts: Alfred C. Welch.....	192

Radvanyi's Public Opinion Measurement. A Survey: Alfred C. Welch.....	416
Rappaport's Diagnostic Psychological Testing: Howard T. Hunt...	662
Rogers and Wallen's Counseling With Returned Servicemen: Donald E. Super.....	565
Sachs' Freud: Master and Friend: K. W. Oberlin.....	293
Scheinfeld's Women and Men: John E. Anderson.....	415
Science Research Associates' Practical Handbook for Counselors: L. E. Drake.....	289
Smith's Handbook of Industrial Psychology: Clifford E. Jurgensen.	288
Smith, Lasswell, and Casey's Propaganda, Communication, and Public Opinion: Donald G. Paterson.....	660
Steiner's Where Do People Take Their Troubles?: Richard M. Elliott	412
Wells and Ruesch's Mental Examiner's Handbook: Paul E. Meehl..	293

Miscellaneous

Erratum.....	668
New Books, Monographs, and Pamphlets...116, 196, 295, 418, 573, 667	

Journal of Applied Psychology

EDITED BY: DONALD G. PATERSON, UNIVERSITY OF MINNESOTA

Consulting Editors

PAUL S. ACHILLES, *Psychological Corporation*; WALTER V. BINGHAM, *A.G.O., War Department*; HAROLD E. BURTT, *Ohio State University*; ARTHUR I. GATES, *T. C. Columbia University*; JOHN G. JENKINS, *University of Maryland*; IRVING LORGE, *T. C. Columbia University*; QUINN MCNEMAR, *Stanford University*; WILLARD C. OLSON, *University of Michigan*; JAMES P. PORTER, *Swarthmore, Pennsylvania*; EDWARD K. STRONG, JR., *Stanford University*; MORRIS S. VITELES, *University of Pennsylvania*; JOSEPH ZUBIN, *N. Y. Psychiatric Institute*.

Table of Contents

<i>The Psychological Corporation's Index of Public Opinion:</i> H. C. LINK	1
<i>Color Aptitude Test, 1940 Experimental Edition:</i> F. L. DIMMICK	10
<i>Statistical Analysis of an Industrial Rating Chart:</i> D. J. BOLANOVICH	23
<i>Simplified Form for Reporting Test Results:</i> B. BAXTER AND E. POTECHIN	32
<i>Comparison of the Reliability and Performance for the Minnesota Rate of Manipulation Test for Subjects Tested Individually and in Groups of Two:</i> J. TUCKMAN	37
<i>The Comparative Validities of Two Tests of General Aptitude in an Army Special Training Center:</i> CAPT. W. D. ALTUS	42
<i>Use of the Shipley-Hartford Test in Evaluating Intellectual Functioning of Neuropsychiatric Patients:</i> LT. M. E. WRIGHT	45
<i>Social I. E. Scale for the Minnesota Multiphasis Personality Inventory:</i> LEWIS E. DRAKE	51
<i>Interests of Senior and Junior Public Administrators:</i> E. K. STRONG, JR. . . .	55
<i>Comparison of the Thurstone and Likert Techniques of Attitude Scale Construction:</i> A. L. EDWARDS AND K. C. KENNEY	72
<i>The Economy of Item Analysis with the IBM Graphic Item Counter:</i> LT. W. J. MCNAMARA AND LT. E. WEITZMAN	84
<i>The Interrelationship of Visual Acuity at Different Distances:</i> W. J. GIESE . .	91
<i>Book Reviews</i>	107
<i>New Books, Monographs, and Phamphlets</i>	116

Published Bi-monthly by The American Psychological Association, Inc.
With the Cooperation of The American Association for Applied Psychology
Prince and Lemon Sts., Lancaster, Pa., and Northwestern University, Evanston, Illinois

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the Act of March 3, 1879
Copyright, 1945, by The American Psychological Association, Inc.



Journal of Applied Psychology

Vol. 30, No. 1

February, 1946

The Psychological Corporation's Index of Public Opinion

Henry C. Link

The Psychological Corporation, New York City

This is the thirteenth in this series of surveys begun in 1937 as experimental studies in social psychology. It was made during the first three weeks in October with 5,000 personal interviews in 123 cities and towns representing a cross-section of the urban and small town population. The interviews were made by 459 interviewers under the direction of 130 psychologists. Two questionnaires were used, each with one-half the sample, so that some questions were asked of 5,000 people and others of only 2,500. The number of interviews for each question is given in the tables. All interviews were made in the home, but only one in a family. Half were made with women, half with men.

The interviews were distributed by four socio-economic groups referred to in the following tables as A, B, C, and D. This distribution was made in accordance with the socio-economic maps in each locality according to which the local supervising psychologist assigned the calls to be made by streets and blocks. The great differences between thinking of these various socio-economic groups are shown in some of the tables. These differences, incidentally, are also an indication of the thoroughness with which these interviews have been distributed by socio-economic levels.

Some of the questions in this survey were asked for the first time. Others are questions which have been asked in as many as seven or eight earlier studies. Where questions have been repeated, the results of a few of the earlier studies are included. Some of these questions were asked also of 1,000 college students in colleges throughout the country. The results are included in the tables.

The Trend Toward Socialism

The greatest and most obvious trend throughout the world today is the trend toward complete Government control of business, variously referred to as Socialism, Collectivism, Totalitarianism, Dictatorship, Fascism, or Communism. Does the American public think that this

country is also headed for Socialism? Are they ready for Socialism? Here are a few straws in the wind.

Q. "Did you know that the English people elected the Socialist-Labor party last July?" (Asked of half the sample, or 2,500.)

Answers	Total	Socio-Econ. Groups				Coll. Studs.
		A	B	C	D	
	%	%	%	%	%	%
Yes.....	72	92	84	70	49	86
No.....	28	8	16	30	51	14
Total Interviews.....	2500	250	750	1000	500	1000

This table illustrates the extreme differences in socio-economic groups on the matter of knowledge. Whereas 8 per cent in the A group said they did not know about the English elections, 51 per cent of the D group professed ignorance. The results of the next question also show interesting differences but in a quite different direction.

Q. "Do you think that this means that the U. S. is also going toward Socialism?"

Answers	Total	Socio-Econ. Groups				Coll. Studs.
		A	B	C	D	
	%	%	%	%	%	%
Yes.....	27	35	32	26	19	28
No.....	53	56	57	53	46	62
Don't know.....	20	9	11	21	35	10
Total Interviews.....	2500	250	750	1000	500	1000

This table shows differences between socio-economic groups on a matter of opinion. According to these results, the higher the socio-economic status, the higher the belief that the United States is or is not headed for Socialism. This strange result may represent a conflict between wishful and realistic thinking. Our studies in other connections have shown that a great many people, especially in the C and D groups, representing the large population of industrial workers, do not know what Socialism is, and do not know what Capitalism is either.

The following two questions were asked in the reverse order in one-half the interviews. There was no significant difference in the results, and, therefore, the results have been combined.

Q. "Is it good for America in peacetime for the Government to set top prices which stores and factories may charge for their goods?"

Answers	Total	Socio-Econ. Groups				Coll. Studs.
		A	B	C	D	
	%	%	%	%	%	%
Yes.....	51	36	46	53	61	46
No.....	43	61	49	41	29	51
Don't know.....	6	3	5	6	10	3
Total Interviews.....	5000	500	1500	2000	1000	1000

Q. "Is it good for America in peacetime for the Government to set top limits for workers' wages and salaries?"

Answers	Total	Socio-Econ. Groups				Coll. Studs.
		A	B	C	D	
	%	%	%	%	%	%
Yes.....	39	28	34	41	51	38
No.....	52	67	59	51	36	60
Don't know.....	9	5	7	8	13	2
Total Interviews.....	5000	500	1500	2000	1000	1000

Here the variations between socio-economic groups are not only considerable, but consistent. The higher the socio-economic group, the higher the rejection of price and wage controls, and vice versa. Regardless of status, a substantial proportion of the urban population approves wage controls and, even more, price controls. Obviously, insofar as these controls are imposed by the Government on industry, Socialism is substituted for a free economy.

Take-Home Pay and Shorter Hours

That the American public is not completely unrealistic in its thinking is shown by the answers to the following questions:

Q. "If a man was paid \$50 a week for 48 hours work in wartime and he is now working only 40 hours a week, should he still be paid \$50?"

Answers	Total	Socio-Econ. Groups				Coll. Studs.
		A	B	C	D	
	%	%	%	%	%	%
Yes.....	42	29	38	45	52	30
No.....	50	65	56	47	39	66
Don't know.....	8	6	6	8	9	4
Total Interviews.....	2500	500	750	1000	500	1000

These results represent the effects of considerable wishful thinking. It would have been interesting, from the research point of view, to see what these answers would have been if we had used the sum, \$80, instead of \$50 in our question. We should also have asked: "Do you think that such increases can be made generally without a corresponding increase in prices?"

Take-Home Pay and the Cost of Living

In view of the controversy over the effects of reconversion on wages in reference to the cost of living, the testimony of people themselves as to their present status is unusually interesting. Because this question was asked in a number of earlier studies proper comparisons of the present with the past can be made. The question asked was as follows:

Q. "Is your family more prosperous (or better off) today than two years ago, less prosperous, or the same?"

Answers	Oct. 1941 %	Oct. 1943 %	Apr. 1945 %	Oct. 1945 %	Coll. Studs. %
More prosperous.....	38	29	28	32	31
About the same.....	47	46	48	51	55
Less prosperous.....	15	23	21	15	12
Don't know.....	—	2	3	2	2
Total Interviews.....	2000	2500	2500	2500	1000

Therefore, in spite of the abrupt termination of many war industries, and the wholesale changeover from wartime to peacetime jobs, an even greater majority of people, 83 per cent, claim that they are as prosperous or more prosperous than they were two years ago. Moreover, this prosperity extends pretty evenly through all socio-economic groups as shown in the following table.

Oct. 1945	Total %	Socio-Econ. Groups			
		A %	B %	C %	D %
More prosperous.....	32	32	31	29	39
About the same.....	51	50	53	53	47
Less prosperous.....	15	16	15	15	12
Don't know.....	2	2	1	3	2
Total Interviews.....	2500	250	750	1000	500

The Public's Spending Plans

The current high prosperity of the public is further demonstrated by the replies to the question:

Q. "How much of the money you have saved since the war do you expect to use within a year or two after the war stops—all of it, two-thirds of it, one-third of it, or none of it?"

Of all respondents, 21 per cent said that they were uncertain, while 8 per cent said they had no savings. The remainder, as compared with those who answered the same question in October 1944, answered as follows:

Oct. 1944 %	Oct. 1945 %	
48	56	planned not to spend any savings
24	21	planned to spend one-third
15	9	planned to spend two-thirds
13	14	planned to spend all their savings

This and the next question were asked after we had first talked with people about their buying plans and after they had enumerated the things they were planning to buy. After they had stated their buying intentions, we asked:

Q. "Do you intend to pay for these things out of your current earnings, or by using the cash you have in the bank, or by cashing in your war bonds, or by buying them on the installment plan?"

Answers	%
Current earnings.....	45
Cash in bank.....	32
Installment plan.....	23
War bonds.....	8
Don't know.....	10
Total Per Cent.....	118*

* Per cents add to more than 100 because some people gave two or more answers.

The Returning Veterans

The rapidity with which service men and women are returning and adjusting themselves to civilian life and peace jobs is indicated by the answers to the following questions:

Q. "How many members of your family (*living at home*) have been in the armed services? How many have come home for good? How many of these now have jobs?"

	No.	%
Families who had 1 or more members in the armed services.....	1004	40
None in the armed services.....	1496	60
Total Interviews.....	2500	

These results show how extensively demands of war affected the families of the United States, at least the urban population. The 1,004, or 40 per cent, of families who had one or more members in the armed services had a total of 1,514 members in the armed services. Of this number 459, or *30 per cent, have come home for good*. Of those who have come home for good, *65 per cent have jobs*.

Who is Doing the Best Job of Reconversion

In view of the sharp controversies over the reconversion from war jobs to peace jobs, the following question was thought timely and was asked in half our sample, or 2,500 families:

Q. "How do you think that the changeover from war jobs to peace jobs is being handled up to now? For instance, do you think the Government has done a good, fair, or poor job? How about the Army—good, fair, or poor? The Navy? The Labor Unions? Employers or businessmen?"

Agencies	Good %	Fair %	Poor %	D.K. %
The Government.....	35	33	20	12
The Army.....	50	23	13	14
The Navy.....	49	21	10	20
The Labor Unions.....	16	16	54	14
Employers or businessmen.....	40	30	14	16

Evidently, the Army and Navy rate very well in the eyes of the public in spite of the many criticisms that have been levelled against them for holding up demobilization. The Government and employers rate about equally well, but the Labor Unions are rated as having done a poor job by a conspicuous majority.

In the other half of our sample we repeated a question which has been used periodically since October 1941, and which represents reconversion in its broadest phases. This question was:

Q. "Who do you think can do the best job in straightening things out after the war: the Government in Washington; Business Leaders; Labor Union Leaders; or others?"

Answers	Oct. 1941 %	Oct. 1943 %	Apr. 1945 %	Oct. 1945 %	Coll. Studs. %
Gov't in Washington.....	47	42	46	51	62
Business Leaders.....	26	28	21	22	21
Labor Union Leaders.....	5	8	8	9	4
All three together.....	7	9	11	12	10
Others or had no opinion.....	17	17	17	11	7
Total Interviews.....	2000	2500	2500	2500	1000

The above per cents add up to more than 100 because some people named two agencies. Reliance on the Government is at a high point. The results by socio-economic groups are especially significant, as may be seen from the following table:

Answers	Socio-Econ. Groups			
	A %	B %	C %	D %
Gov't in Washington.....	43	46	55	56
Business Leaders.....	36	29	17	12
Labor Union Leaders.....	3	6	10	15
All three together.....	19	15	12	7
Others or had no opinion.....	7	10	10	15
Total Interviews.....	250	750	1000	500

Unemployment Compensation

In view of the controversy over the desirability of extending the periods and amounts of unemployment compensation, the following question was asked:

Q. "Do you think that unemployment insurance is keeping many people who have been laid off from taking new peacetime jobs?"

Answers	Total %	Socio-Econ. Groups				Coll. Studs. %
		A %	B %	C %	D %	
Yes.....	52	65	60	51	39	43
No.....	36	26	32	38	42	49
Don't know.....	12	9	8	11	19	8
Total Interviews.....	2500	250	750	1000	500	1000

Although sharp differences are shown by socio-economic groups, a large proportion of people even in the large C and D industrial groups say that they believe unemployment insurance at present is keeping many people from taking new peacetime jobs.

Public Postwar Optimism

The sharp swing toward greater optimism in postwar prospects which was shown by the April 1945 survey was maintained in the present survey and, in some respects, even heightened. This is particularly true in respect to wages. The questions and the results were as follows:

Q. "During the next year or two do you think that the people of this country will be better or worse off than they are now?"

Answers	Oct. 1941 %	Oct. 1943 %	Apr. 1944 %	Apr. 1945 %	Oct. 1945 %	Coll. Studs. %
Will be better off.....	13	48	31	55	49	42
Will be worse off.....	69	32	47	27	32	47
Don't know.....	18	20	22	18	19	11
Total Interviews.....	2000	2500	2500	2500	2500	1000

Q. "How about jobs; will there be more, fewer, about the same?"

Answers	Oct. 1941 %	Oct. 1943 %	Apr. 1944 %	Apr. 1945 %	Oct. 1945 %	Coll. Studs. %
Will be more jobs.....	8	26	22	34	35	21
About the same.....	11	20	17	22	26	28
Will be fewer jobs.....	74	46	51	38	33	49
Don't know.....	7	8	10	6	6	2

Q. "Will wages be higher, about the same, or lower?"

Answers	Oct. 1941 %	Oct. 1943 %	Apr. 1944 %	Apr. 1945 %	Oct. 1945 %	Coll. Studs. %
Will be higher.....	10	8	6	10	25	11
Will be about same.....	20	26	26	35	34	32
Will be lower.....	60	60	60	51	36	55
Don't know.....	10	6	8	4	5	2

Q. "Will our Government be less democratic, more democratic, or the same as now?"

Answers	Oct. 1941 %	Oct. 1943 %	Apr. 1944 %	Apr. 1945 %	Oct. 1945 %	Coll. Studs. %
Less democratic.....	26	19	17	22	15	19
Same as now.....	33	34	31	37	45	45
More democratic.....	19	30	27	28	27	30
Don't know.....	22	17	25	13	13	6

The Public Predicts the Next War

In view of the tremendous interest in peace and measures for a permanent peace, we repeated a question which we asked first in a depth study in February 1943 (Link, H. C., An experiment in depth interviewing on the issue of internationalism vs. isolationism, *Pub. Opin. Quart.*, 1943, 6, 267-279). The question was as follows:

Q. "After this war, do you think that we will make a peace settlement that will last, or do you think that we will have another world war in twenty-five years or so?"

Answers	Feb. 1943 %	Oct. 1944 %	Apr. 1945 %	Oct. 1945 %	Coll. Studs. %
Will have another war.....	43	54	51	59	71
Will make a lasting peace.....	47	28	33	28	22
Don't know.....	10	18	16	13	7
Total Interviews.....	200	2500	2500	2500	1000

Q. "Who do you think will be our next enemy?"

Answers by Those Who Said There Would be Another War

	Oct. 1944 %	Apr. 1945 %	Oct. 1945 %		Oct. 1944 %	Apr. 1945 %	Oct. 1945 %
Russia.....	29	27	37	England.....	4	4	3
Germany.....	9	6	2	China.....	1	1	1
Japan.....	5	3	5	Don't know....	6	10	11
Total.....					54	51	59

This reflects a steady and sharp increase in the percent who expect another war, except for the April 1945 period which reflected the result of the San Francisco Conference. There is a sharp increase in those who believe that the next war will be with Russia. Of the 59 per cent who anticipate another war, about 70 per cent name Russia as the next foe.

Travel by Airplane

In view of the tremendous development of aviation through the war and its bright prospects after the war, we repeated a question asked in 1943, namely:

Q. "Have you ever taken a trip in an airplane? (If Yes) How many?"

Answers	Oct. 1943 %	Oct. 1945 %
Yes.....	28	31
No.....	72	69
Total Interviews.....	2500	2500

We then asked the following question in the October 1945 study to obtain some measure of the extent to which some people were planning to travel by air in the near future, namely:

Q. "Are you planning to take such a trip in the near future, say within a year?"

Answers	Oct. 1945 %
Yes	23
No	67
Don't know	10
Total Interviews	2500

Since this question was not asked before, a comparison is not possible. Nevertheless, it indicates that a very large number of people are planning to travel by air shortly.

Received November 23, 1945.

References to Published Studies *

1. Corby, P. G., Roslow, S., and Wulfeck, W. H. Consumer and opinion research: Experimental studies on the form of question. *J. appl. Psychol.*, 1940, 24, 334-346.
2. Link, H. C. How many interviews are necessary for results of a certain accuracy? *J. appl. Psychol.*, 1937, 21, 1-17.
3. The Psychological Corporation. A study of public relations and social attitudes. *J. appl. Psychol.*, 1937, 21, 589-602.
4. Link, H. C. A study of public opinion and morale. *J. appl. Psychol.*, 1941, 25, 636-645.
5. Link, H. C., and Freiberg, A. D. The problem of validity vs. reliability in public opinion polls. *Pub. Opin. Quart.*, 1942, 5, 87-98.
6. Link, H. C. Workers' reactions to industrial problems in a war economy. *J. appl. Psychol.*, 1942, 26, 416-438.
7. Link, H. C. The eighth nation-wide social experimental survey. *J. appl. Psychol.*, 1943, 27, 1-11.
8. Link, H. C. An experiment in depth interviewing on the issue of internationalism vs. isolationism. *Pub. Opin. Quart.*, 1943, 6, 267-279.
9. Link, H. C. The ninth nation-wide social experimental survey. *J. appl. Psychol.*, 1944, 28, 1-15.
10. Link, H. C. The tenth nation-wide social experimental survey. *J. appl. Psychol.*, 1944, 28, 363-375.
11. Link, H. C. The eleventh nation-wide social experimental survey. *J. appl. Psychol.*, 1945, 29, 103-107.

* A complete set of references to studies published by The Psychological Corporation is available on request.

A Color Aptitude Test, 1940 Experimental Edition

Forrest Lee Dimmick *

Hobart College

For a long time there has been a need in many industries for a means of determining the suitability of workers for their jobs in fields which require the matching of colors. Expressions of this need have appeared as a rule in house organs or in reports to technical associations.¹ One attempt at least has been made to deal directly with the problem, but while it pointed the way, its materials are not available for general use.²

At the annual meeting of the Inter-Society Color Council in February 1939, in a session given over to discussion, the problem was formulated and a call made for assistance. Dyers of silk textiles, for example, are given a swatch of cloth or a skein of yarn which must be duplicated. They must decide the dyes to be used and the precise amount of each component that is required, by successive trial dyeings. Hence, dyers must make quick and accurate judgments of color. While extreme deficiencies in color vision are not likely to go undetected for long, it is obvious that an industry does not wish to train a new worker in the use of dyes only to find that he cannot discriminate between them. On the other hand anomalies less than "color blindness" cause serious difficulties because they go unnoted until some glaring error is committed. An example of such a case is a dyer who, in a series of dyeings of a particular red, made every successive match slightly too yellow with the result that the final dyeings had to be rejected. Of equal importance is a sheer ineptness with colors that may appear because aptitude for making matches is often the very last consideration that is taken into account in selecting apprentice dyers.

As a result of its discussion, the Color Council directed its Problem Committee to undertake the development of a test or a series of tests that would answer the needs of various industries and of other fields where color matching plays an important role.³

* Co-Chairman, Inter-Society Color Council Committee on Problem 10, Color Aptitude Test.

¹ American Pulp and Paper Association. *Color blindness*, Report No. 34, 1941.

² W. O. D. Pierce. *The selection of color workers*, London, Pitman, 1934.

³ In order to make clear why such a problem could be undertaken with some confidence, the organization of the Inter-Society Color Council should be understood. The Council consists primarily of the representatives of thirteen national organizations,

When we attempted to formulate a test procedure, it was evident that some kind of matching technique was indicated, in which a degree of manipulation plays a part. Two methods came up for final consideration: 1. The arrangement of finely graded series in proper order was pro-

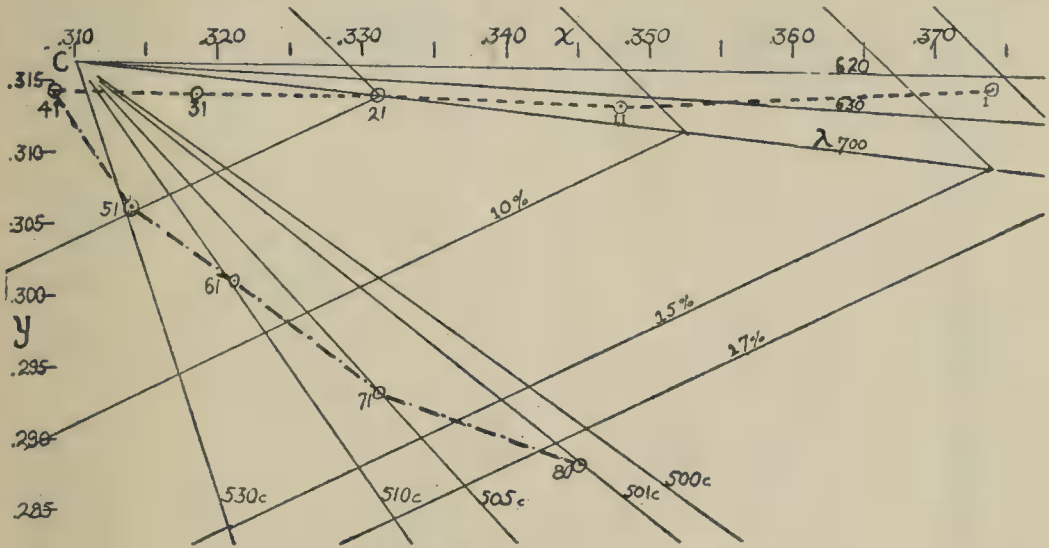


FIG. 1. Section of the I.C.I. diagram showing specifications of the Aptitude Test material. The dotted line represents the yellowish-red series, the dot-dash line the bluish-red series.

posed since it called for fine discrimination; and 2, the matching of individual color samples suggested itself as psychologically basic.

The projected test needed to be something more than an evaluation of color discrimination in the sense of a measure of some retinal capacity.

namely: American Artists Professional League; Amer. Asso. of Textile Chemists and Colorists; Am. Ceramic Soc.; Am. Psychological Assoc.; Am. Soc. for Testing Materials; Federation of Paint and Varnish Production Clubs; Illuminating Engin. Soc.; Am. Pharmaceutical Assoc.; Optical Soc. of Am.; Soc. of Motion Pict. Engin.; Tech. Asso. of the Pulp and Paper Industry; Textile Color Card Assoc. of the U. S.; U. S. Pharmacopoeial Convention. Not only do these groups bring to the Council their numerous problems in color, but they offer a wealth of technical and practical information. Thus, the color test upon which we are working has required a technical skill in preparation of materials that would have been prohibitively expensive on an individual basis, if available at all. In addition, diverse industries in which color matching plays an important part are open for standardization purposes, on populations in which relatively high color skills have been demonstrated. The committee appointed by the I.S.C.C. to develop the test is constituted as follows: Co-Chairmen, Forrest L. Dimmick, A.P.A., and Carl E. Foss, A.S.T.M.; I. A. Balinkin, A.C.S.; C. Z. Draves, A.A.T.C.C.; W. C. Granville, O.S.A.; J. P. Guilford, A.P.A.; Le Grand Hardy, I.M.; H. Helson, A.P.A.; D. B. Judd, O.S.A.; N. Macbeth, I.E.S.; Elsie Murray, A.P.A.; S. M. Newhall, A.P.A.; D. Nickerson, O.S.A.; J. L. Parsons, T.A.P.P.I.; L. Sloan, A.P.A.; A. H. Taylor, I.E.S.; M. J. Zigler, A.P.A.

“Color aptitude” involves the utilization of discriminative ability with greater or less efficiency. From the beginning, we realized that the factor of color-blindness must be taken into account, though it is only a minor part of our problem. In order to deal, therefore, with color aptitude and color blindness in the same test, we decided to work with colors that have

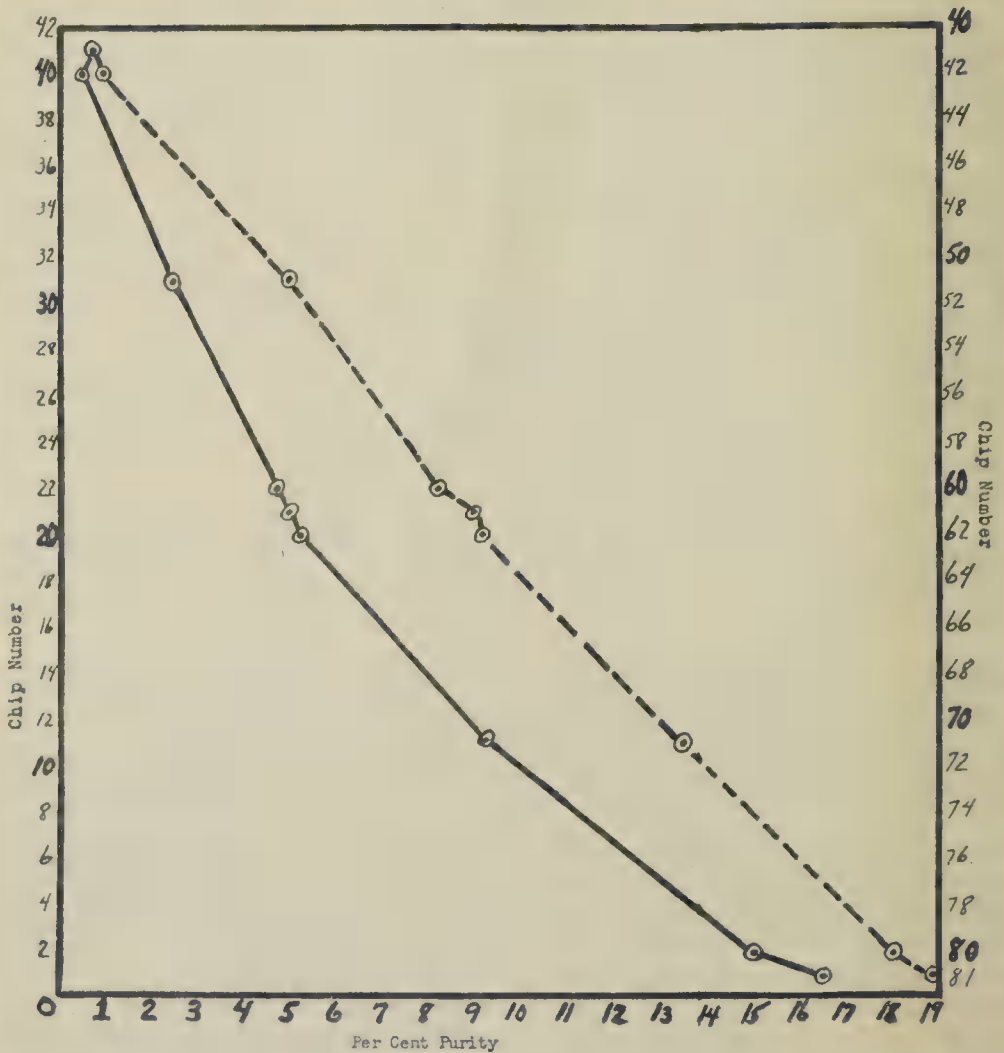


FIG. 2. Per cent colorimetric purity of the two series. The solid line represents the yellowish-red series, 1-40, the dotted line, the bluish-red series, 41-80.

been shown to lie in the neutral zones for so called “deuteranopia” and “protanopia.” We chose for the initial tests two reds, a bluish red (Munsell 6 RP/5 ($\lambda 500c$)) and a yellowish red (Munsell 5 R/5.4 ($\lambda 612$)). We proposed to make a 40-step saturation series from each of these colors to neutral gray (Munsell N/5). It was much easier to propose such materials than to produce them. The Munsell Color Company had

found it practicable to make up their Book of Color with swatches varying by two units each, while we were proposing to divide every one of these units into 10 steps, i.e. 20 to their one. Mr. Foss and Mr. Granville who undertook the project for the committee accomplished the job with remarkable success.

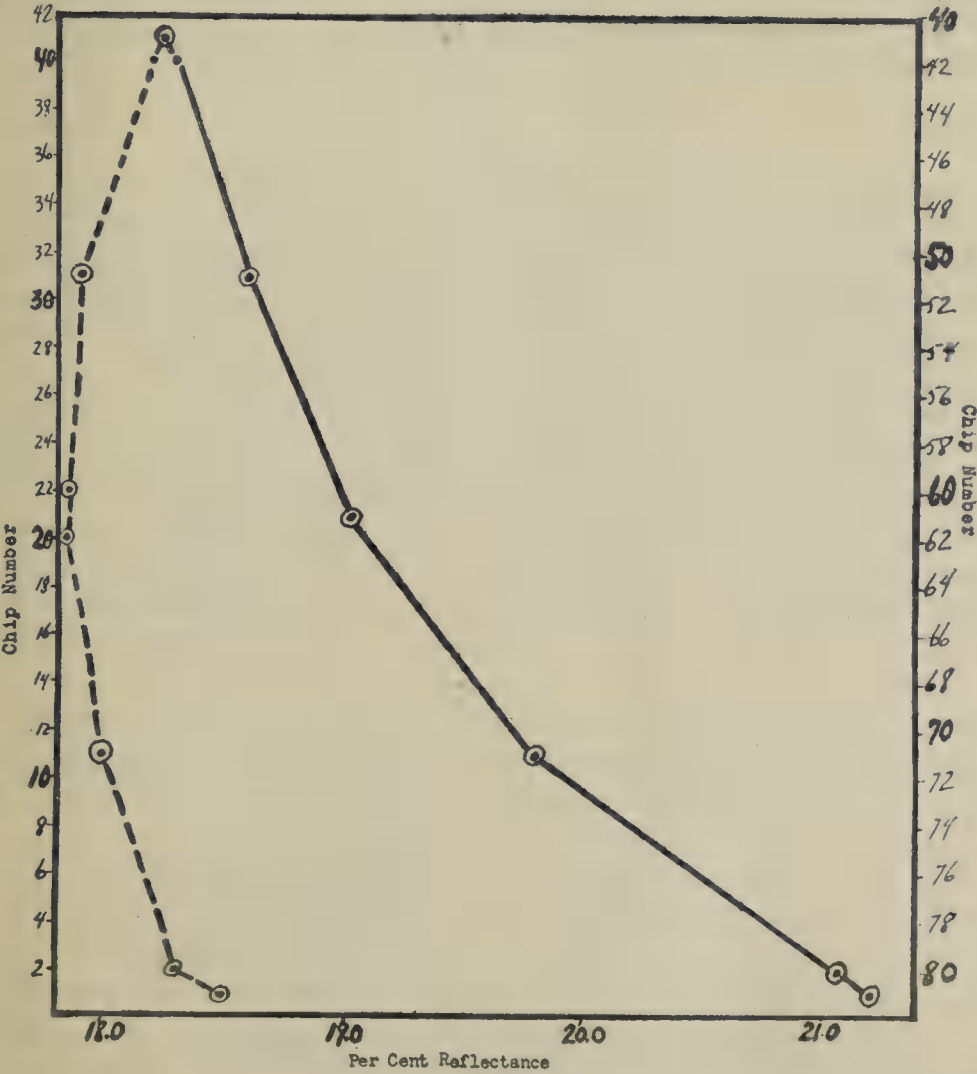


FIG. 3. Per cent reflectance of the two series of chips. The solid line represents the yellowish-red series, 1-40, the dotted line, the bluish-red series, 41-80.

Figure 1 shows the location of the color series on the I.C.I. chromaticity diagram. From it can be seen the degree to which they parallel lines of constant dominant wave length. The maximum deviation in the yellow-red series is effectively in the region of $3m\mu$, and in the purple-red series $10 m\mu$. Figure 2 gives the change in per cent purity within the

series. Both series approximate straight lines thus indicating that we have nearly equal steps of colorimetric purity. Figure 3 is a similar plot of the per cent reflectance of the chips in both series. That of the bluish-red series shows constant reflectance plus or minus an extreme of 0.3 per cent. The yellowish-red series which was chosen to lie in the neutral zone for protantopes was made to increase by 3 per cent in reflectance from the neutral point to its saturated end to compensate for the darkening of the spectrum at the red end for such subjects.

Procedure

The first procedure to be tried consisted of presenting the subject with the 80 chips in a haphazard arrangement on a gray background under daylight illumination of 50 foot candles and requiring him to put the series in correct order. Surprisingly, the problem proved much too easy. A number of subjects made the arrangement without error on the first trial, therefore, while we could use it to screen out markedly deficient subjects, it gave no differentiation among color competent subjects. The failure of this procedure is psychologically interesting because it points out that "discrimination" is not a single, simple concept.⁴ In some psychophysical experiments which we carried out with the series of chips, we found that for most 0's the J. N. D.⁵ lies between the first and second chips away from any given standard. Apparently then, the serial arrangement includes factors in addition to "discrimination" of each chip from its two neighbors. A casual examination of the perceptual problem indicates that an inversion in the series reveals itself to a subject on the basis of at least 3 discrimination judgments some of which are between definitely supraliminal stimuli. Thus the inversion of a pair of chips gives the perception of a "hump" in the series.

Our next procedure consisted of laying out one set of chips in a predetermined haphazard order and having the subject match a second set to the first. The task is somewhat more difficult than the first procedure. Scores in terms of errors in matching spread out over a greater range, but it remained possible for some subjects to make perfect sets of matches. This was due in some measure to the gradual elimination of choices as successive chips were filled in and to the fact that the subjects, necessarily, were permitted to rearrange the matches until they obtained a satisfactory total result.

The final procedure is to require a subject to find in the haphazardly

⁴ K. Koffka, *Perception: An introduction to the Gestalt-Theorie* Psychol. Bull., 1922, 19, 540 ff.

⁵ F. B. Titchener, *Experimental psychology*. New York: Macmillan, 1927, Vol. II, part 1, p. 59.

arranged field a match for a single chip at a time. When found, a match is recorded and the chip is laid aside, so that the range of choices remains constant throughout the test. In spite of the fact that it was not possible to control several factors in the administration of the Preliminary Sets, no subject made a perfect score on his first trial. With this procedure we obtained 65 scores based upon 80 judgments. Analysis of them gives us indications of several revisions that must be made. Figure 4 shows a frequency distribution of the 65 error scores. From it we concluded no more than that the scores are adequately distributed for our purposes and that the curve promises to follow a normal form.

A major difficulty was encountered by many of the committee members who tried out the test, in the time required to complete 80 matches. The average time per match was 1.16 minutes, which gives an average total time for 80 matches of 92.8 minutes. Obviously, an hour and a half is too long for this sort of test. Many subjects could not be got to do the full 80 judgments in one sitting, nor could they return easily for a second session. This is why three times as many cases were obtained, in which only one-half the test was completed. We reexamined the data from the 65 cases for an indication of how we might interpret the half-tests of only 40 judgments. We correlated the whole scores with part scores with the results shown in Table 1:

Table 1
Correlations

	<i>r</i>	<i>PE</i>
80 jud. with 40 judg. (1 to 20) + (61 to 80)	.88	.016
80 jud. with 20 judg. (1 to 10) + (71 to 80)	.75	.036
10 jud. with 10 judg. (1 to 10) + (71 to 80)	.53	.06

The intercorrelations could have been made in many different ways, but we carried them no further since these gave sufficient indication that we could reduce the length of the test without materially altering the distribution of the results.

A fractionated distribution of errors as shown in Table 2 gives further indication of the fundamental homogeneity of the test. Therefore, we divided the "Preliminary Sets" into two sets each; one including chips 1-10, 21-30, 41-50, 61-70 and the other the remaining 40 chips. The standard tests consist, then, of duplicate sets of 40 chips each, one permanently mounted on a neutral gray background and the other unmounted.

The gray background forms the inside of a shallow wooden box 12 by 10 by 1 inches. The chips themselves are $1\frac{1}{2}$ by $1\frac{1}{4}$ inches. These dimensions allow us to fasten them in 4 rows of 5 chips in each of the two halves of the background with $\frac{1}{2}$ inch between chips in the rows. Subjects must match a loose chip to a fixed one by placing it below the latter. This is accomplished both by instruction and by having a $\frac{1}{2}$ in. strip of gray cardboard along the top edge of every row. Other spatial factors are kept constant by using the same "haphazard" arrangement of the fixed array. This "haphazard" array is not entirely chance. The yel-

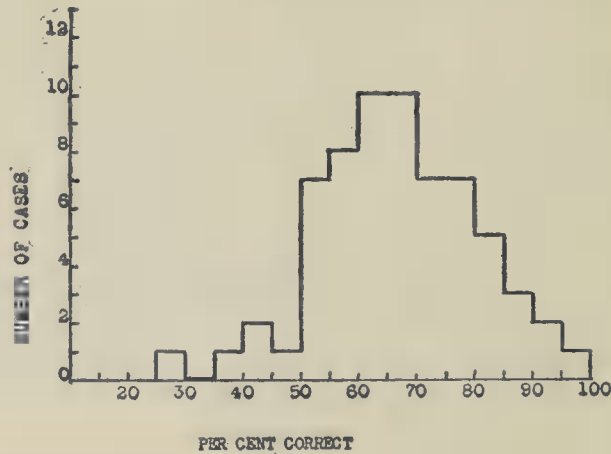


FIG. 4. Distribution of 65 scores based on 80 matching judgments with unlimited time.

lowish-red and bluish-red chips are alternated both horizontally and vertically, and care is taken that near matching chips are well scattered over the field.

The time factor presented several further problems which required other modifications of procedure. While the reduction of the total number of judgments from 80 to 40 brought the average time for the test within acceptable limits, individual times scattered widely about that average. Speed of matching may be as important a factor in color aptitude as accuracy. We sought to take speed into account by weighting

Table 2
Fractionated Distribution of Errors

	Yellowish-red				Bluish-red			
Chip nos.	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
Errors	447	424	372	336	322	337	477	457
Totals	1579				1587			

scores with a factor obtained from the ratio of the individual time to the average time.

$$\text{Weighted score} = \text{per cent correct} + \frac{(\text{Av. time} - \text{ind. time})}{\text{av. time}} \times \frac{\text{Per cent correct}}{2}$$

$$Sw = \% + \frac{(Ta - Ti) \%}{Ta} \cdot \frac{1}{2}$$

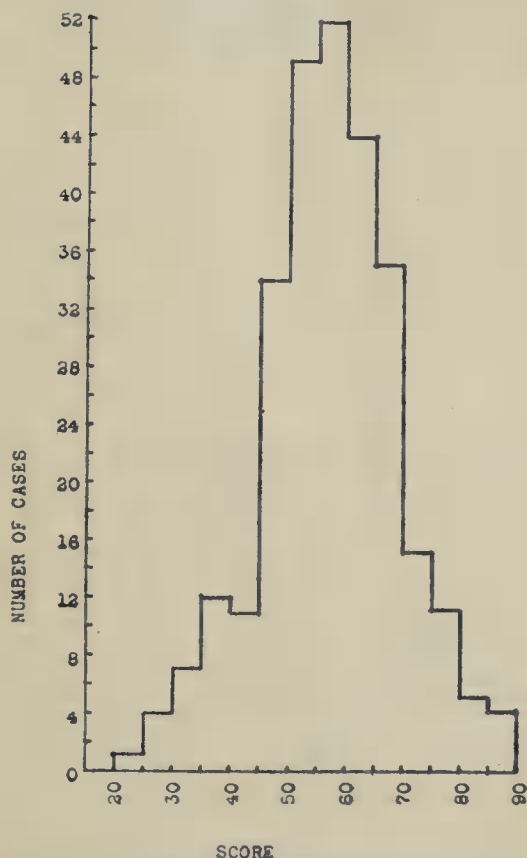


FIG. 5. Distribution of 338 scores using the final procedure with 30 min. time limit.

Thus the weighted scores became a measure of accuracy in matching modified by the rate of matching.

This device, however, did not bring the actual time for administering the test within desired limits nor stabilize it at a specific time and was given up in favor of a fixed time of 30 minutes, which is sufficiently below the unrestricted average time (45 min.) so that the majority of subjects may be expected not to complete the test. Fixing a time limit introduces the possibility of an effect by the order in which the chips are presented for matching. Some matches may be easier to make than others. When

all 40 chips were presented, this factor could be neglected, but now the particular chips matched within the time limit may be an important determinant of the score. Therefore, the loose chips are presented in a specified order so that all subjects match the same chips up to any point in the course of the test.

Complete instructions for giving the test are as follows:

"The set of 40 mounted chips which constitutes the Matching Field should be laid out flat in diffuse illumination of *daylight quality* equivalent to I.C.I. 'Illuminant C.' The illumination should be in the neighborhood of 50 foot candles and as uniform as possible over the whole area of the 'matching field.' Care must be taken that the illuminant does not shine into the subject's eyes

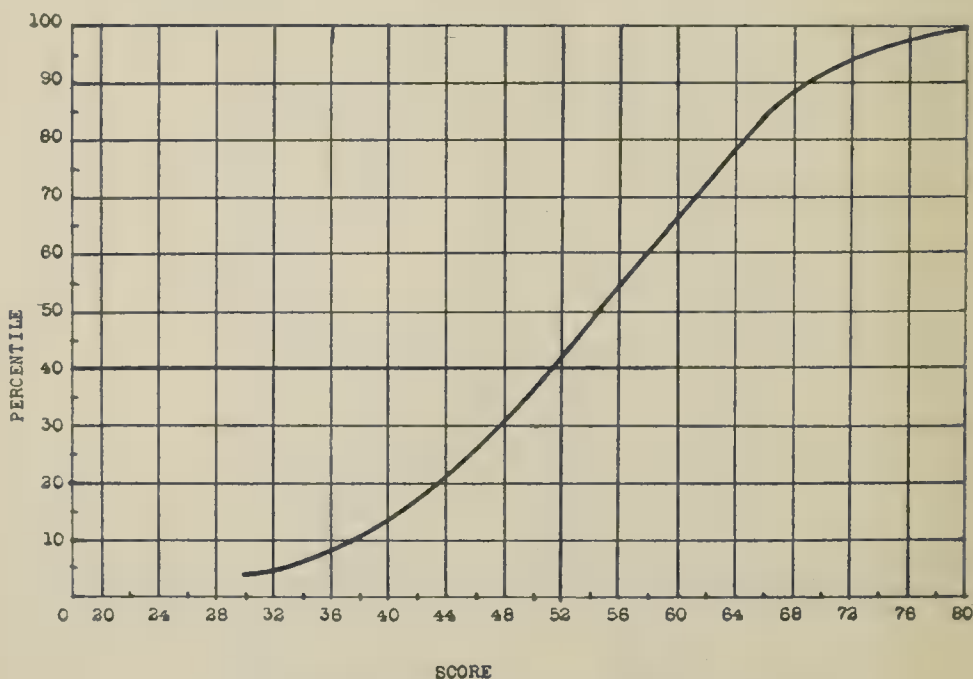


FIG. 6. Percentile distribution of 338 scores.

and that no shadows are cast upon the 'matching field.' Images mirrored by the glossy surfaces of the chips can be eliminated by proper placing of them with relation to the illuminant and the subject.

"Procedure: Give the subject the first one of the unmounted chips to be matched and instruct him as follows:

"Among the colored chips laid out before you, there is a mounted one which matches exactly each one of the unmounted set. Find the mounted chip which matches each unmounted one. Do not hurry, but work right along, because *your score will be determined partly by your speed in matching.* You will be allowed to make as many matches as you can in 30 minutes. Your score will be best if you use all the time allowed.

"When the subject has found the match, enter the code letters of the chip in the corresponding place on the Scoring Chart, put the first chip back in the box and give the subject the second chip. *Note that spaces are left on the Scoring Chart for recording several unmounted matches to every mounted chip,*

but the code symbol of an unmounted chip may be recorded *only once*. There will be some blank spaces on the completed Scoring Chart.

"The subject may move a 'matching chip' about in order to compare it with any chip in the 'matching field.' (1) He must place it *below* the field chip with which he is comparing it, and in contact with it. (2) All matches must be made with both chips flat on the surface of the background. (3) The subject must view the chips from a distance of *not less than* 10 inches. (4) The subject matches only one chip at a time."

We now have 338 test scores obtained with the final test procedure. The subjects include expert colorists, textile students, industrial workers, office workers and college students. Figure 5 shows the distribution of scores and Figure 6 their percentile distribution. The curves approxi-

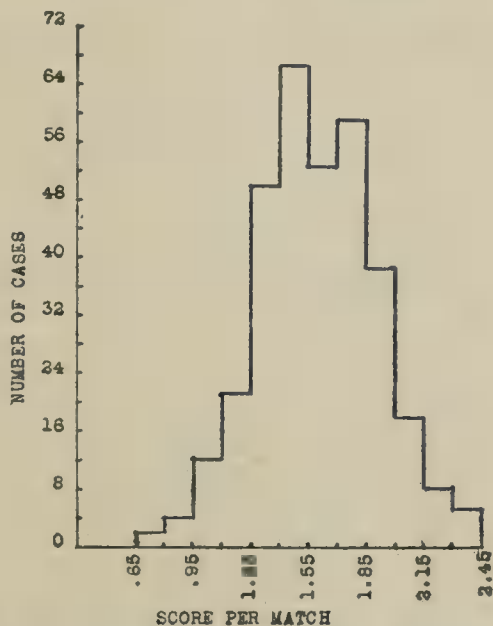


FIG. 7. Distribution of the accuracy factor, score per match, obtained by dividing the score made in 30 min. by the number of matches; 338 scores.

mate the normal curves nearly enough to indicate the general character of our sample population. An attempt was made to obtain other information concerning the color proficiency of subjects, with which to correlate test scores. A rating scale of nine steps from "Exceptional" to "Poor" "Expertness in Color Matching" was printed in every score sheet, but it was rarely completed because the information was lacking or not readily available to the person who was giving the test. It appears that these highly desirable ratings can not be obtained on a general basis. This, indeed, is the reason for the test. It should be possible, however, to obtain them within more limited groups, such as a particular industry or a single laboratory or company. Even then a correlation

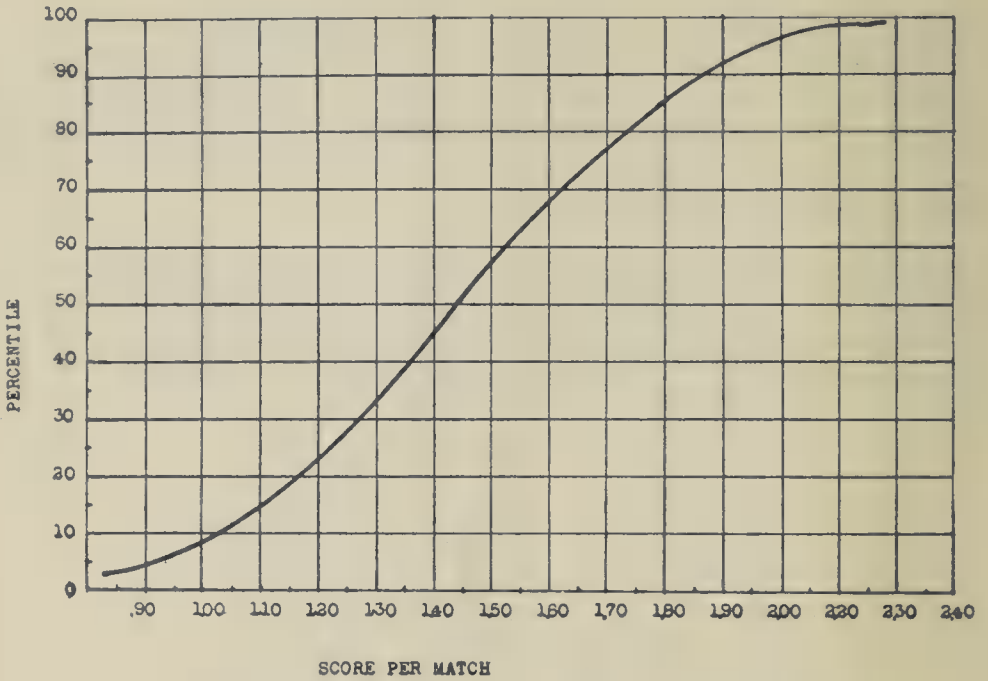


FIG. 8. Percentile distribution of the accuracy factor.

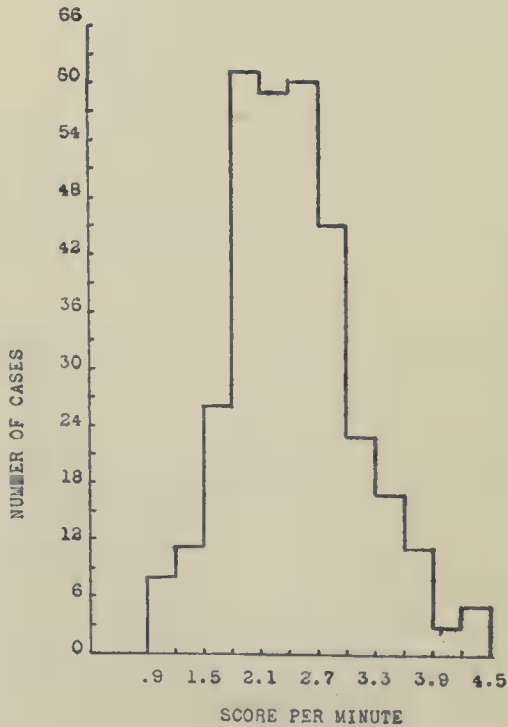


FIG. 9. Distribution of the speed factor, score per minute, obtained by dividing the score made in 30 min. by 30 min. (or by the actual time when the test was completed in less than 30 min.). $N = 338$ cases.

will measure the judgment technique and the validity of individual judges.

In the discussion of the procedure, we pointed out the desirability of combining speed and accuracy in a single score. For particular problems, it is interesting to know the contribution made by each of these factors. In order to separate them, we calculated the *score per match* and the *score per minute* for every subject. The distribution and percentile curves for the two factors are shown in Figures 7-10. Correlation be-

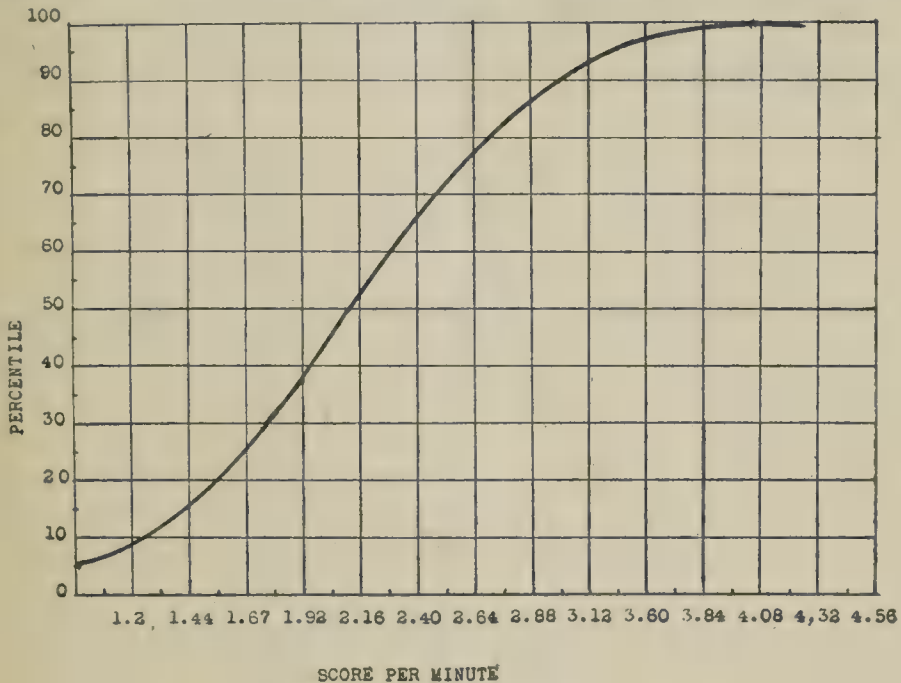


FIG. 10. Percentile distribution of the speed factor.

tween them is positive, though not high, viz. $r = +.35$, $PE_r = .045$, indicating that there is only a slight correlation and therefore one cannot be substituted for the other.

Summary

On the basis of the foregoing results we have found that matching judgments made within saturation series of finely graded steps give a well distributed set of individual scores upon which to establish "color matching aptitude" ratings. While as yet these ratings are related only hypothetically to practical performance on the job, there is agreement among those who have used the test that the results are significant. In the few cases where supervisors would give a definite rating to the color competence of a subject, the correlation was good. Within one company

which tested a large number of its employees, chemists, and dye workers, rated higher than laboratory technicians and office workers.

The materials used in the first experimental test have been expended in this preliminary work, therefore the committee has produced new material for a "1944 Experimental Edition" of the test which is now available in a limited quantity. Since the major criticisms of the first form of the test was its limited range of hues, and its relatively low saturation, the new edition corrects these faults. No data have yet been obtained from its use, but the sets are now in the hands of the committee and results will be published as soon as they become available.

Received January 3, 1945.

Statistical Analysis of an Industrial Rating Chart *

D. J. Bolanovich

Radio Corporation of America

Wherever extensive and continued use is made of a rating device, its value can be enhanced by thorough statistical analyses. This fact is, of course, very obvious and well-known to all who have had even slight training in the use of ratings. However in the industrial field, it is not unusual to find rating devices in use which are accepted without question as adequate measures of everything they specify. It is not unusual to find preassigned scoring methods whose suitability has never been tested. Too seldom do those applying the ratings attempt to discover what the scales really measure, and to recognize overlapping, bias, and those changes in intended scale values which are imposed unknowingly by raters themselves. When men are selected for promotion by such rating reports, superficial interpretations may do an injustice to employees and employers. At best the ratings are not as efficient as they might be.

This report describes the experience of one company, which uses a Personnel Rating Chart, and which endeavors to obtain maximum effectiveness of its interpretation through statistical analysis. The Personnel Rating Charts were developed to meet the company's need for records upon which to base promotion of field personnel to key managerial and technical positions. The subjects of ratings used in this study were 143 field engineers who service electronic equipment throughout the United States. Raters were 11 district managers supervising the engineers. It is the company's policy to rate all personnel other than higher management every six months. The foregoing analysis was made of the first semi-annual ratings.

The confidential nature of information on the charts prevents their illustration here. However, a brief description of the charts may make this discussion more understandable. Items composing the charts were: (1) Personality, (2) Personal Appearance, (3) Punctuality, (4) Thoroughness, (5) Efficiency, (6) Resourcefulness, (7) Dependability, (8) Coopera-

* This rating experiment was based on the experience of the R.C.A. Service Company, a subsidiary of the R.C.A. Victor Division of the Radio Corporation of America. Acknowledgment is made to Mr. A. Goodman, assistant general manager of the company, who made the study possible, and Mrs. E. Fish, who tabulated the data.

tion, (9) Job Attitude, (10) Technical Ability, (11) Sales Ability, (12) Organizing Ability, (13) Judgment, and (14) Desire for Self-Improvement. These items were deemed most important by company management for promotion to supervisory and specialist positions. Each item was carefully defined in terms of observable behavior. Each item was followed by five boxes numbered from 1 to 5, and representing continuous intervals on a scale. The meanings and limits of the intervals were indicated at the top of the chart. An attempt to curb bias was made by changing the order of the numbered boxes from item to item. An example of two such items would be:

1. <i>Personality</i> : Cheerfulness and pleasantness in relations with others. Extent of friendships with associates and customers. Ability to hold confidence and admiration.	<table><tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr></table>	1	2	3	4	5
1	2	3	4	5		
2. <i>Resourcefulness</i> : Success in handling routine and special problems without continual help. Activity in developing new applications, finding new needs for equipment. Suggestions for equipment, methods, procedures, and making new products.	<table><tr><td>4</td><td>5</td><td>1</td><td>2</td><td>3</td></tr></table>	4	5	1	2	3
4	5	1	2	3		

Where the key to the scales reads: 5 = Excellent (among the best 10%); 4 = Superior (among the best 1/3, but not in best 10%); 3 = Good (in the middle 1/3); 2 = Fair (in the lower 1/3, but not in lowest 10%); and 1 = Poor (among low 10%).

An additional item asked, "Would you recommend this employee for a more responsible assignment?" and requested specific information. Other information was asked for on the reverse side of the sheet. Each supervisor making ratings was given an explanatory chart for reference which discusses the purposes and uses of the personnel ratings, and contains helps for effective rating.

After completion of one set of ratings, duplicate copies of the charts were forwarded to the home office, where all ratings were punched on IBM cards for analysis. Distributions of ratings for each item, and intercorrelations between each pair of items were calculated on IBM sorters. Correlations were also determined between "recommendation for more responsible assignment" and each item. These latter correlations were biserial. Inter-item correlations were Pearsonian. The data thus obtained were processed as follows:

- (a) A factor analysis was made to determine the extent to which common factors accounted for the variance in ratings.
- (b) Using "Yes" and "No" responses to the question, "Would you recommend this man for a more responsible assignment?" as a criterion, a multiple correlation was obtained between the best combination of items and the criterion.
- (c) Methods of scoring the charts were experimented with to find a statistically sound scoring procedure.

Factor Analysis

Factor analysis of the items was conducted according to Thurstone's Centroid Method.¹ The extraction of factors was continued until McNemar's criterion² was satisfied. McNemar proposes that when the standard deviation of the partial residuals reaches or falls below $\frac{1}{\sqrt{N}}$, the magnitude of residuals may be considered as due to chance sampling errors in the original intercorrelations. The present extraction of factors was at first halted after five were obtained, at which time the S. D. of partial residuals was only .039 greater than $\frac{1}{\sqrt{N}}$. However, when rotation failed to reduce the items loadings on some of the factors satisfactorily, a sixth factor was extracted. Inclusion of the sixth factor improved the effectiveness of rotation in reducing factor loadings to zero and eliminating negative loadings.

Tables 1 and 2 show factor loadings, communalities, and uniqueness of items before and after 15 completed rotations. It appeared that any further rotations would not be effective in clarifying the factors further. Rotations were made using two axes at a time and computing new loadings after each rotation.³ All six factors had some item loadings greater than .40. The factor F_1 contained 7 items whose loadings were not significantly greater than zero; F_2 contained 3 such; F_3 contained 9; F_4 contained 3; F_5 contained 7; and F_6 contained 4. Except for the factor F_4 , not many items were heavily loaded with any given factor. This made interpretation of the factor meanings not too difficult a job.

In attempting names for the factors, the definitions given for the items were carefully considered. Factor F_1 , for example, has heaviest loadings in Personal Appearance and Thoroughness, which seem difficult to reconcile. However, on the chart Personal Appearance is defined as "Careful-

¹ Guilford, J. P., *Psychometric methods*. New York: McGraw-Hill Book Co., 1936, pp. 478-508.

² McNemar, Q., On the number of factors. *Psychometrika*, 1942, 7, 9-18.

³ Guilford, J. P., *op. cit.*, p. 502.

Table 1
Factor Loadings and Communalities for the Fourteen Ratings
Before Rotation of Axes

Item	F ₁	F ₂	F ₃	F ₄	F ₅	F ₆	h ²
1. Personality	.729	-.377	-.144	.213	.210	.155	.81
2. Personal Appearance	.460	-.437	-.438	.034	.167	-.172	.65
3. Punctuality	.654	.078	-.186	.226	-.386	-.021	.67
4. Thoroughness	.733	.259	-.306	-.186	-.152	-.044	.76
5. Efficiency	.808	.207	-.102	-.184	.008	-.158	.76
6. Resourcefulness	.730	.130	-.114	-.201	.301	.124	.70
7. Dependability	.826	.247	-.088	.136	-.129	.135	.80
8. Cooperation	.797	-.197	.170	.161	.059	.112	.74
9. Attitude Toward His Job	.767	-.232	.218	-.069	-.148	.124	.73
10. Technical Ability	.668	.318	.222	-.279	.244	.212	.78
11. Sales Ability	.559	-.274	.162	-.260	-.239	-.163	.56
12. Organizing Ability	.762	.133	.128	.084	.000	-.201	.66
13. Judgment	.766	.040	.255	.167	.189	-.178	.75
14. Desire for Self-Improvement	.596	.125	.161	.220	-.076	.199	.49

ness in dress and posture." It seemed that F₁, then, was a "Meticulousness" or "Attendance to Detail" factor. F₂ which is highly loaded in Technical Ability, Resourcefulness, and Efficiency, would seem to be similar to the factor reported by Ewart, Seashore, and Tiffin ⁴ as "Ability to do present job." F₃, with highest loadings in Sales Ability and Attitude Toward Job, was termed a "Sales Ability" factor. Selling is one important aspect of the work of field engineers. F₄ is difficult to name since many items measure it. It would seem that a factor of "Job Conscientiousness" runs through those items with large loadings. F₅ was termed an "organizing" or "systematic" factor since it was found largely in the items Judgment and Organizing Ability. Judgment is defined on the charts as "ability to make good decisions based on facts." F₆ seems to be a "Social Intelligence" factor. Its heaviest loadings are in Personality, Cooperation, Judgment, and Desire for Self-Improvement.

These names for factors, of course, are subjective "best guesses." It might be interesting to note a simple check made on the appropriateness of factor names. A local university psychology instructor wrote down the six suggested factor names. Then without knowledge of factor loadings, he wrote his opinion as to whether an item would receive a high, low, or intermediate loading in each factor. His guesses seemed to approximate actual figures rather closely. He experienced most difficulty with factor F₅ which was then termed "Executive Ability."

⁴ Ewart, E., Seashore, S. E., and Tiffin, J., A factor analysis of an industrial merit rating scale. *J. appl. Psychol.*, 1941, 25, 481-486.

Table 2
Factor Loadings, Communalities and Uniqueness After Fifteen
Completed Rotations

Item	F ₁	F ₂	F ₃	F ₄	F ₅	F ₆	<i>h</i> ²	<i>u</i> ²
1. Personality	.380	.226	.370	.137	.230	.634	.80	.20
2. Personal Appearance	.572	-.016	.339	.077	.378	.261	.66	.34
3. Punctuality	.334	-.001	.058	.683	.086	.275	.66	.34
4. Thoroughness	.498	.389	.020	.568	.147	.005	.74	.26
5. Efficiency	.341	.501	.030	.508	.355	.088	.76	.24
6. Resourcefulness	.360	.664	.083	.206	.183	.237	.71	.29
7. Dependability	.318	.380	-.026	.636	.061	.387	.80	.20
8. Cooperation	.086	.331	.297	.377	.253	.575	.74	.26
9. Attitude Toward His Job	.016	.343	.461	.491	.190	.349	.73	.27
10. Technical Ability	.016	.812	-.002	.287	.078	.179	.78	.22
11. Sales Ability	-.004	.190	.476	.427	.343	-.010	.56	.44
12. Organizing Ability	.107	.339	-.017	.513	.419	.305	.66	.34
13. Judgment	.006	.389	.001	.364	.493	.473	.75	.25
14. Desire for Self-Improvement	.018	.276	.011	.454	.003	.455	.49	.51

In addition to these common factors, the items Desire for Self-Improvement and Sales Ability seem to have high uniqueness values and may represent specific characteristics. In view of the fact that reliabilities cannot be determined for these ratings, Specificity is not obtainable. However, an attempt was made to estimate roughly the Specificities in the following manner: According to Thurstone's formulae,⁵ the reliability of an item will be at least as great as its calculated communality. Thus, the reliability of the item Personality would be at least .80. From the formula for specificity, $S^2 = r - h^2$, we would get a specificity of zero for Personality. Then, it was assumed that since other items are equally or more objective than Personality, they would have reliabilities of at least .80 also. By substituting .80 for reliability in the above formula for each item, specificities were estimated. This method is only roughly approximate, but gives some idea of the magnitude of specificity values. The items with largest estimated specificities are given in Table 3. It appears that Desire for Self-Improvement and perhaps Sales Ability do measure independent characteristics.

It might be well to point out here some comparisons between this study and that done by Ewart, Seashore, and Tiffin, since the latter has influenced greatly the thinking of those who use rating scales. While the present study found six common factors, that of Ewart, Seashore, and Tiffin found three (one of which was discarded as unreliable). The traits rated differed somewhat, but were similar in many respects. The

⁵ Guilford, J. P., *op. cit.*, p. 477.

chart used by the three authors measured 12 traits. Perhaps the most reasonable explanation for the difference in results of these two studies lies in the nature of the work being rated. The earlier study dealt with direct production workers, while this study deals with field engineers whose work requires a wide range of abilities and characteristics. For example, field engineers sell, they contact people, they make reports, meet many new problems and require highly specialized training. With factory workers, supervisors are largely influenced by quantity and quality of work done, and perhaps by attitudes of employees. There are some further reasons that may account for the greater number of factors in the charts for field engineers: (1) The district managers are highly interested in developing the all-round potentialities of their employees,

Table 3
Estimated Specificities* of Items Having Smallest Communalities

Item	Estimated Specificity
Desire for Self-Improvement	.31
Sales Ability	.24
Organizing Ability	.14
Punctuality	.14
Personal Appearance	.14
Resourcefulness	.09

* Estimated on the basis of reliabilities assumed equal to the communality for the Personality item.

(2) The managers had been accustomed to reporting on various traits of field men, (3) Efforts were made to break up possible bias or halo effects, as can be seen in the above discussion of the nature of the chart used.

Relationship of Items to Overall Performance

Table 4 shows the correlations between individual item ratings and responses to the question, "Would you recommend this man for a more responsible assignment?" It is assumed here that "Yes" and "No" represent a dichotomous division of a continuum of degree of recommendation. Furthermore, to be a practical criterion, recommendations must represent present job success and not peculiar fitness for some specific job. Since it is a general practice for supervisors to recommend for promotion those workers performing best on their present jobs, this criterion should be sufficiently valid.

The highest single correlation (biserial) with the criterion was .59, which was shown for each of the five items: Personality, Efficiency, Thoroughness, Job Attitude, and Organizing Ability. The lowest was

Table 4

Biserial Correlations of Item Ratings With "Yes" or "No" Responses to "Would You Recommend This Employee for a More Responsible Position?"

Item	r_{bis}
1. Personality	.59
2. Personal Appearance	.43
3. Punctuality	.33
4. Thoroughness	.59
5. Efficiency	.59
6. Resourcefulness	.58
7. Dependability	.53
8. Cooperation	.39
9. Attitude Toward His Job	.59
10. Technical Ability	.46
11. Sales Ability	.49
12. Organizing Ability	.59
13. Judgment	.48
14. Desire for Self-Improvement	.24

Table 5

Matrix of Item Intercorrelations*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Personality	.70													
2. Personal Appearance	.60	.60												
3. Punctuality	.42	.29	.68											
4. Thoroughness	.46	.27	.58	.73										
5. Efficiency	.48	.41	.48	.73	.73									
6. Resourcefulness	.49	.38	.36	.58	.64	.67								
7. Dependability	.53	.27	.68	.70	.63	.63	.70							
8. Cooperation	.70	.36	.49	.46	.53	.56	.62	.70						
9. Attitude	.61	.33	.48	.46	.54	.48	.62	.67	.67					
10. Technical Ability	.39	.05	.26	.51	.62	.67	.57	.48	.51	.67				
11. Sales Ability	.33	.30	.37	.38	.43	.33	.38	.48	.58	.31	.58			
12. Organizing Ability	.49	.26	.49	.59	.60	.51	.67	.57	.56	.52	.44	.68		
13. Judgment	.54	.29	.44	.43	.63	.52	.61	.69	.59	.58	.38	.68	.69	
14. Desire for Self-Improvement	.44	.12	.43	.34	.51	.37	.53	.54	.46	.44	.22	.45	.48	.54

* Item self-correlations (Reliabilities) are given as equal to the highest intercorrelation in the row and column in which the item appears.

.24 between the criterion and Desire for Self-Improvement. A maximum multiple correlation of .81 was obtained between the criterion and a combination of the following 7 items: Personality, Efficiency, Resourcefulness, Cooperation, Job Attitude, Sales Ability, and Organizing Ability.

Here the Wherry-Dolittle technique ⁶ was used to select items and was continued until an additional item would have added only .007 to the multiple correlation coefficient. The regression equation for estimating the degree of recommendation for a more responsible assignment is:

$$X_c = 3.8 (\text{Personality}) + 1.0 (\text{Efficiency}) + 3.8 (\text{Resourcefulness}) \\ - 4.8 (\text{Cooperation}) + 2.1 (\text{Job Attitude}) + 2.3 (\text{Sales Ability}) \\ + 2.5 (\text{Organizing Ability}).$$

Method of Scoring

Several possibilities were considered for scoring the charts. First, they could be scored by simply adding the ratings for each scale. Second, a total score could be obtained by weighting each item in proportion to its regression equation coefficient. Third, weights might be assigned items which would yield scores for each of the factors found by factor analysis.⁷ It was decided that factor scores would not be of much practical value.

In order to test the relative effectiveness of scoring by straight item summing and by using regression weights, charts for 94 field engineers were scored both ways. Biserial correlations were then calculated between total scores and recommendations for more responsible assignments. As might be expected from the multiple correlation coefficient, total scores by the regression equation method correlated .83 with recommendations. Total scores obtained by adding all 14 item ratings correlated .75 with recommendations.

As a result of these findings, it was decided to score the charts by assigning regression weights.

Summary and Conclusions

1. Fourteen-item rating charts for 143 field engineers, rated by 11 district managers, were analyzed to determine: (a) the common and unique factors operating, (b) the multiple correlation between rating items and overall job success, and (c) possible methods resulting from these analyses for scoring and interpreting the ratings.

2. Factor analysis showed six common factors measured by the scales. These were named: Attendance to Detail, Ability to do the Present Job, Sales Ability, Conscientiousness, Organizing or Systematic tendency, and Social Intelligence.

3. The items Sales Ability and Desire for Self-Improvement had

⁶ *The Wherry-Dolittle test selection method* (courtesy of R. J. Wherry). Work-guide used by U. S. Employment Service Division, Occupational Analysis Section.

⁷ Thomson, G. H., *The factorial analysis of human ability*. New York: Houghton Mifflin Co., 1939, pp. 107-110.

large uniqueness values and are probably measuring significant specific factors.

4. When a multiple regression equation was computed for predicting recommendations for promotion, 7 of the 14 items were included as adding significantly to the multiple correlation. The multiple R was .81, and included the items: Personality, Efficiency, Resourcefulness, Cooperation, Job Attitude, Sales Ability, and Organizing Ability.

5. Total scores obtained by adding the 14 item ratings yielded a correlation of .75 with recommendations for more responsible assignments. Total scores obtained by summing the products of item scores and their regression weights for 7 items yielded a correlation of .83 with recommendations. The latter method was selected for assigning overall scores.

6. These findings and their application will aid the company in its use of the scales in the following ways:

- (a) An objective method is made available for immediate determination of relative all-round performance of employees.
- (b) Relationships of various items to job success give some idea of their relative importance.
- (c) In considering individuals, factor analysis findings give a better idea of the meanings of item scores.

Received December 6, 1944.

A Simplified Form for Reporting Test Results *

Brent Baxter

Owens-Corning Fiberglas Corporation, Toledo, Ohio

and

Evelyn Potechin

Ohio State University

The success of a personnel testing program in an industrial or government agency depends to a large extent on the efficiency with which test information is conveyed to the operating officials such as supervisors, employment interviewers, counselors, and training instructors. These individuals are usually untrained in psychological testing and are unable to understand readily any technical aspects of a testing program or to understand the significance of test scores as applied to personnel problems. If they are to derive any benefit from the use of test results, the test technician must present the results to them in an easily understood manner. This paper describes a "Test Report Form" that has been found effective in expressing a test score or a set of test scores for an individual to non-psychologically trained operating officials.

Several devices are in current use for translating the raw scores of tests to something more meaningful. Among these are the standard score, the percentile, descriptive phrases (e.g., excellent, good, average, poor, very poor), and the profile. The Test Report Form (Fig. 1) is a combination of all these devices. It is essentially a graphic rating scale on which the check marks are positioned according to percentiles or standard scores based on objective test scores.

Spaces are provided at the top of the Form for the examinee's name and other identifying data. Beneath this is a brief explanation to the operating official to guide him in understanding the Form. In the left-hand column are listed by name all tests which are in use in the program. Brief phrases further describing what the test measures may be placed beneath the name of each test. The list of tests shown in Figure 1 was adjusted to include all those which were in the regular available series

* The authors wish to express their appreciation to Dr. Charles C. Gibbons whose suggestion prompted the development of the Form, and to Mr. Roger T. Lennon who offered several useful criticisms during the construction of the Form and in reading the manuscript.

for clerical workers in a particular government agency and was designed for use with other materials which explained the purpose of each test. Since in most cases an individual does not take all tests in the list, the names of the tests administered are encircled in red for ready identification.

Opposite the name of each test is a graphic scale, along which five phrases descriptive of various levels of performance are placed in order. The test results are recorded on these scales in terms of percentiles, using the small dots on the lines as guides to indicate the deciles. A testing clerk may record the results by noting the individual's percentile and placing a red check mark on the line at the point which conforms to the proper percentile. The supervisor interprets the test results directly in terms of the descriptive phrases and does not have to have the concept of percentiles explained to him. For those who wish to know the percentile score, it may be read fairly accurately from the position of the red check mark. The same procedure can be adapted to the use of standard scores.

On the present Form the extreme phrases each refer to about seven percent of the norm group, the second and fourth phrases each pertain to 23 per cent, and the center phrase concerns the middle forty per cent. The decile dots have been spaced along the lines so that equal spaces along each line refer to equal increments in ability. The total line length represents five standard deviations (plus and minus two and one-half standard deviations).¹

No attempt is made to join the red check marks representing performance on the various tests as on a profile chart or psychograph. The relative standing on the various tests is clear without a joining line, which would only be confusing in this situation; a line joining check marks for non-adjacent test scales will cross the scale for another test which may not have been administered and suggest that scores have been obtained for the intermediate test. It is important, moreover, that if an individ-

¹ The following method may be used to obtain this system of spacing: (1) determine the length of line most convenient for the size of paper to be used (6 inches is suitable for paper 8 inches wide); (2) determine the length represented by a standard deviation by dividing the line length by 5; (3) place the 5th decile dot (mean) in the center of the line; (4) place the 6, 7, 8, and 9th decile dots to the right of the mean by .25, .52, .84, and 1.28 of the standard deviation length (obtained in 2) respectively; (5) place the 4th, 3rd, 2nd, and 1st decile dots to the left of the mean by .25, .52, .84, and 1.28 of the standard deviation length respectively. In arranging the descriptive phrases, the extreme phrases at the right and left are placed beyond plus and minus 1.47 of the standard deviation length respectively. The second phrase is placed between the end phrase and the 3rd decile dot; the middle phrase is between the 3rd and 7th decile dots; and the fourth phrase is between the 7th decile dot and the extreme phrase (+1.47 sigma). The spaces allotted to the phrases are only approximately equal.

TEST
REPORT

NAME		DATE	
AGE	SEX	EDUCATION	GRADE AND TITLE
SECTION		BRANCH	

EXPLANATION: In the column on the left is a list of tests; those taken by the individual whose name appears above are encircled in red. Opposite the name of each test is a series of phrases arranged in order along a line to describe increasing levels of test performance. Test results are indicated by a red check mark on the line above the phrase which best describes the individual; sometimes the mark is more accurately placed on the line between two phrases.

TESTS	DESCRIPTIVE PHRASES					
Learning Ability	Learns Very Slowly	Can Learn Lowgrade Work of Simple, Routine Nature	Can Learn Moderate Difficulty Tasks	Can Learn Fairly Complex Duties	Able to Learn the Most Complex Tasks	
Mechanical Aptitude	Not Recommended For Training in Any Mechanical Work	Can Learn Simple, Routine Mechanical Tasks	Can Learn Mechanical Work of Average Difficulty	Can Learn Fairly Difficult Mechanical Work	Can Learn Most Difficult Mechanical Work	
Clerical Aptitude	Unusually Slow for Clerical Jobs of Any Kind	Suitable for Simple Clerical Work Where Speed is Not Essential	Suitable for Tasks Requiring Average Clerical Ability	Has Superior Speed on Most Clerical Tasks	Can Perform Clerical Tasks With Unusual Efficiency	
Shorthand Taken At _____ WPM	Very Inaccurate	Numerous Mistakes	Average	Number of Errors	Few Mistakes	Practically No Errors

ERRORS		Very Inaccurate	Numerous Mistakes	Average Number of Errors	Few Mistakes	Practically No Errors
Typing	WPM					
	SPEED					
Military Correspondence Information		Very Slow	Slower Than the Average Typist	Types the Average Number of Words Per Minute	Superior Typing Speed	Exceedingly Rapid Typist
		Practically No Information	Little Knowledge	Average Mastery of Military Correspondence Rules	Well Informed	Exceedingly Well Informed
		Low	Below Average	Average	Above Average	Excellent
		Low	Below Average	Average	Above Average	Excellent
		Low	Below Average	Average	Above Average	Excellent

COMMENTS:

Fig. 1. Sample copy for Test Report Form.

ual's standings on the various tests are to be compared, the groups on which the tests' percentiles are based should be comparable.

At the bottom of the Form is a space for "Comments." This is used for making recommendations regarding hiring, transferring, upgrading, etc. Any unusual results or "highlights" of the individual's performance are also discussed here.

The Test Report Form cannot be used, however, as a substitute for an explanation of tests and their uses. A careful explanation of the advantage and limitations of test results in general as well as of the specific tests used should be given to any non-technician who receives the Test Report Form as part of the basis for any personnel action.

Some difficulty may be encountered in devising descriptive phrases for each of the tests but no more so than in any graphic rating scale. Care must also be exercised in seeing that where ever "absolute" descriptive phrases (e.g., can learn fairly complex duties) are used beneath the relative percentile scale, the absolute terms have a valid meaning. For an extreme example of this kind of error, if the typing test norms were based on a group of first-class typists, even typists in the 3rd percentile could not be said to be very slow or very inaccurate. If the results of only one or two tests are to be presented, it is possible to arrange a form whose vertical axis gives a description of absolute performance and whose horizontal axis shows the standing in a group. This two-axis form, however, becomes rather complicated and more difficult to explain.

Summary

Some advantages of using the Test Report Form include the following:

1. Standard interpretations of the test results are recorded for each test.
2. Once the Form has been arranged, the marking of the test interpretation is very easy and can be done by a testing clerk.
3. No statistical knowledge is required on the part of the operating official who uses the test information. It is easily explained to any "reader".
4. It avoids any system which separates the total range of scores into a limited number (e.g., five) of groups. The concept of continuity of performance becomes more apparent.
5. The Form may be used as the testing unit's permanent test record if the raw score is placed beneath the name of the test.

Received December 13, 1944.

A Comparison of the Reliability and Performance for the Minnesota Rate of Manipulation Test for Subjects Tested Individually and in Groups of Two

Jacob Tuckman

Jewish Vocational Service, Cleveland, Ohio

The Minnesota Rate of Manipulation Test ¹ is used to select workers for office and factory jobs where speed of hand and finger manipulation is important. The apparatus is a wood board containing 60 cylindrical holes, arranged in four rows of fifteen into which 60 slightly smaller blocks can be placed. The test consists of two parts: Placing, in which the subject, using one hand, places the blocks into the holes in a definite order from a fixed position; and Turning, in which the subject picks up a block with the left hand, turns it over, and puts it back into the same hole with the right hand, alternating hands for each subsequent row. One practice and four test trials are given. The score is the number of seconds required to complete the four test trials.

The manual of directions accompanying the test states "It is advisable to have at least two people taking the test at the same time. The competing effect will stimulate each into doing his best, thus increasing reliability." Since, in practice, it is not always possible to test in groups of two or more, the purpose of this study is to determine whether there are differences in reliability and in test performance between subjects tested individually and those tested in groups of two.

Test scores for Placing and Turning were available for 255 boys and 208 girls tested individually, and 185 boys and 200 girls tested in groups of two. For those tested individually the mean age was 16.0 for both boys and girls; for those tested in groups of two the mean age was 16.3 for boys, and 15.9 for girls. The four groups were superior in intelligence as measured by the ACE Psychological Examination for High School Students and College Freshmen (1939-1942 editions), and the Terman Group Test of Mental Ability. For those tested individually, the mean percentile rank for intelligence was 80 for boys and 76 for girls; for those tested in groups of two, the mean percentile rank was 78 for boys and 75 for girls. All were enrolled in a college preparatory course in several senior high schools or were enrolled in junior high schools normally leading

¹ Developed by W. A. Zeigler and distributed by The Educational Test Bureau, Inc., Minneapolis, Minnesota.

to this course of study. Each of the four groups was about equally distributed in grades 9–12. The median grade for each of the four groups was 10A (latter half of the 10th grade).

The reliability coefficients for Placing and Turning for the four groups are presented in Table 1. The coefficients were obtained by correlating scores on Trials 1 and 3 and scores on Trials 2 and 4, and corrected by applying the Spearman-Brown prophecy formula. For Placing and Turning, the reliability coefficients tend to be higher for both boys and girls tested individually. These data are not in agreement with Zeigler's

Table 1
Split-half (Trials 1 and 2 and Trials 3 and 4) and Corrected Reliability Coefficients for Placing and Turning for High School Boys and Girls Tested Individually and in Groups of Two

	N	Placing			Turning		
		<i>r</i>	P.E. <i>r</i>	Corrected	<i>r</i>	P.E. <i>r</i>	Corrected
				<i>r</i> *			<i>r</i> *
Individually							
High School Boys	255	.88	.0098	.93	.93	.0059	.96
High School Girls	208	.89	.0096	.94	.90	.0092	.95
Combined Group	463	.88	.0068	.94	.92	.0049	.96
In Groups of Two							
High School Boys	185	.87	.0121	.93	.91	.0176	.95
High School Girls	200	.87	.0119	.93	.84	.0200	.91
Combined Group	385	.87	.0085	.93	.88	.0075	.94

* Corrected by Spearman-Brown prophecy formula.

findings. In comparing the combined groups, the difference is greater for Turning than for Placing. The $\frac{D}{P. E. \text{ diff.}}$ is 1.51 in favor of those tested individually for Placing and 3.98 for Turning, but these differences are not statistically reliable.

The mean, standard deviation, and skewness of the scores for Placing and Turning for the four groups are given in Table 2. The comparison of the mean scores for the groups is presented in Table 3.

Although the reliability is not increased, the performance of subjects tested in groups of two is faster than that of subjects tested individually. These differences are statistically significant when the performance of boys or girls tested in groups of two is compared with that of boys or girls tested individually. In comparing the combined groups, the $\frac{D}{\sigma \text{ diff.}}$ is 9.1 for Placing and 5.6 for Turning, in favor of those tested in groups of

Table 2
Mean Score (Time in Seconds), Standard Deviation, and Skewness of the Distribution for Placing and Turning for All Groups. N = 848

	Placing					Turning					
	N	Mean	S.D.	Sk	σsk	$\frac{Sk}{\sigma sk}$	Mean	S.D.	Sk	σsk	$\frac{Sk}{\sigma sk}$
Individually											
High School Boys	255	234.2	20.2	-3.65	1.70	-2.15	186.1	24.1	-6.40	1.86	-3.44
High School Girls	208	234.3	20.7	-1.75	1.80	-.97	182.4	19.9	-3.10	1.78	-1.74
Combined Group	463	234.2	20.3	-2.90	1.24	-2.34	184.2	22.4	-5.00	1.30	-3.85
In Groups of Two											
High School Boys	185	224.6	16.5	-1.75	1.68	-1.04	178.1	18.6	-3.15	1.70	-1.85
High School Girls	200	220.5	17.3	-2.40	1.60	-1.50	175.2	14.9	-1.95	1.37	-1.42
Combined Group	385	222.5	17.1	-.90	1.15	-.80	176.6	16.9	-2.70	1.08	-2.50

two. For those tested individually, the performance of boys and girls is almost identical for Placing; girls are faster for Turning. For those tested in groups of two, girls are faster than boys for Placing and Turning. These sex differences are not significant. In comparing the performance of the combined groups with the norms of the Educational Test Bureau, the mean score for those tested individually is equivalent to the 48th percentile for Placing, and the 63rd percentile ² for Turning. For those tested in groups of two, the mean score is equivalent to the 67th percentile for Placing, and the 74th percentile for Turning.

Table 3
Comparison of the Mean Scores for Placing and Turning for All Groups

	Placing			Turning		
	D	σ diff.	$\frac{D}{\sigma \text{ diff.}}$	D	σ diff.	$\frac{D}{\sigma \text{ diff.}}$
Individually						
Boys and Girls	.1	1.91	.05	3.7	2.05	1.80
In Groups of Two						
Boys and Girls	4.1	1.72	2.38	2.9	1.73	1.68
Boys (Individually) and Boys (In Groups of Two)	9.6	1.75	5.49	8.0	2.04	3.92
Girls (Individually) and Girls (In Groups of Two)	13.8	1.89	7.30	7.2	1.73	4.16
Boys and Girls Combined (Individually) and Boys and Girls Combined (In Groups of Two)	11.7	1.28	9.10	7.6	1.35	5.60

The distributions of each of the four groups show a tendency for the scores to cluster toward the upper end of the scale. For boys tested individually the skewness is significant for Placing and Turning; for the other three groups the skewness is not significant. When the groups are combined the skewness is significant only for Turning for those tested individually.

The noteworthy differences that exist between those tested individually and those tested in groups of two for Placing and Turning warrant the establishment of separate norms for high school students. These are presented in Table 4.

² For discussion regarding the tendency of subjects to perform more rapidly on Turning than on Placing see J. Tuckman, A comparison of norms for the Minnesota Rate of Manipulation Test. *J. appl. Psychol.*, 1944, 28, 121-28.

Table 4

Cleveland Jewish Vocational Service Norms for Placing and Turning for High School Students (Grades 9-12) Tested Individually and in Groups of Two

Per- centile	Placing (Time in Seconds)		Turning (Time in Seconds)	
	Individually N = 463	In Groups of Two N = 385	Individually N = 463	In Groups of Two N = 385
99	189.0	185.9	144.6	145.7
95	203.4	195.1	154.2	153.0
90	210.4	200.6	159.6	157.4
85	213.5	205.4	163.0	160.7
80	216.6	208.6	166.3	162.6
75	219.7	210.9	168.8	164.5
70	222.8	213.3	171.3	166.5
65	225.9	215.4	173.8	168.7
60	228.3	217.5	176.4	171.0
55	230.7	219.5	178.9	173.1
50	233.1	221.5	181.5	175.2
45	235.7	223.5	184.1	177.2
40	238.3	225.7	186.7	179.2
35	241.0	228.2	189.3	181.1
30	243.8	230.8	192.5	183.1
25	246.8	233.6	196.0	185.5
20	249.8	236.4	200.7	188.4
15	255.2	239.9	207.2	192.5
10	261.6	244.1	213.4	198.3
5	270.2	252.3	221.3	206.5
1	285.8	265.7	247.5	226.9

Summary

The reliability of Placing and Turning is not increased when high school students are tested in groups of two, but the performance of these students on both tests is significantly faster than that of students tested individually.

Received December 11, 1944.

The Comparative Validities of Two Tests of General Aptitude in an Army Special Training Center *

William D. Altus, Capt., AGD

Camp McQuaide, California

In a previous article by Bell and Altus,¹ the numerous objectives of an Army Special Training Center have been described. It is sufficient here to say that the main function of such a Center was to bring the trainees to a level in reading, writing and arithmetic which the Army recognized as literate. Literacy, as defined by the Army, may be roughly compared with the achievement of the average public school fourth grader in the tool subjects. The trainee had to reach this level within twelve weeks or be discharged as inapt. Very infrequently a trainee was shipped if he possessed a skill which would make him quite valuable to the Army, even though he was illiterate.

For the first few weeks after the Ninth Service Command Special Training Center was organized (September, 1943), the test of general aptitude administered to the incoming trainees was the Wechsler-Bellevue Intelligence Test. When it was finally possible to obtain a set of the officially sanctioned general ability scale, The Wechsler Mental Ability Scale, Form B, it was immediately put into use. Both tests are much alike, the second deriving from the first mentioned. Both are administered individually.

Since the disposition of the trainee was practically dichotomous (a few were discharged for physical reasons), it was possible to compute validating bi-serial coefficients of correlation for the various tests used by the Personnel Consultants' Section of this Center. The respective validities of certain subtests of the Wechsler Mental Ability Scale, Form B, have been previously reported by Altus.² A recapitulation of the validities

* The opinions expressed in this article are those of the author and are not to be construed as reflecting the official attitude of the Army of the United States. 1st Lt. Ephraim Yohannan, Pfc. Sidney Feinberg, Sgt. Carl Karasek and T/5 Grant Smith are to be credited with tabulating the original data presented in this article. Lt. Yohannan is responsible for the statistical work involved in the study.

¹ Bell, H. M., and Altus, W. D. The work of psychologists in the Ninth Service Command Special Training Center. *Psychol. Bull.*, 1944, **41**, 187-191.

² Altus, W. D. The differential validity and difficulty of certain verbal and performance subtests of the Wechsler Mental Ability Scale. *Psychol. Bull.*, 1945, **42**, 238-249.

for four verbal subtests will be found in Table 1. Also given in the same table are the validities for the five verbal subtests of the Wechsler-Bellevue.

It will be noted that there are over five times as many cases involved in the validating coefficients for the Army Wechsler as for the civilian Wechsler. For that reason, greater confidence can be placed in the bi-serial correlations for the Form B Scale. It is noteworthy that the Arithmetic subtest of the Army version of the Wechsler is the most valid of the four subtests in use at this Center, while the Arithmetic subtest of the Wechsler-Bellevue was the least valid of the five subtests originally administered. The difference between these two coefficients is almost significant ($D/P.E._D$ of 3.58). Apparently the validity of this subtest

Table 1

The Comparative Validities of Certain Verbal Subtests of the Wechsler Mental Ability Scale, Form B, and of the Wechsler-Bellevue Intelligence Scales in Predicting the Disposition of Trainees in an Army Special Training Center

Subtest	Form B			Wechsler-Bellevue		
	r_{bis}	P.E. $_{r_{bis}}$	N	r_{bis}	P.E. $_{r_{bis}}$	N
Arithmetic	.467	.018	1991	.290	.046	367
Information	.406	.018	1991	.475	.044	367
Comprehension	.360	.019	1991	.462	.045	367
Similarities	.334	.020	1991	.323	.046	367
Digit Span*				.442	.045	367
Total Scale	.553	.017	1991	.579	.044	367

* Digit Span was administered to the earlier group (Wechsler-Bellevue) only.

was markedly improved through revision for Army use; perhaps one should rather say that its validity was markedly improved for trainees in an Army Special Training Center.

The Information subtest proved to be quite valuable in both versions of the Wechsler. In Form B, Information takes second place only to the Arithmetic; in the original scales, it was the most valid of the subtests, its validity coefficient being even slightly higher than is Arithmetic for the Army version. Comprehension takes third place for the Army Wechsler; for the Bellevue it takes second. The original version of Comprehension is quite obviously a much better test for the restricted type of mentality found in a Special Training Center, the r_{bis} for the Bellevue being .102 higher. The Similarities subtest is least valid for Form B and has next to the lowest validity in the Bellevue Scales. The validating coefficients for this subtest are about the same, .334 and .323.

In the form of the Army Wechsler employed here, the subtest on Digit Span was not administered. While the Digit Span test is not so valid as Information and Comprehension in the Bellevue, it is markedly better than Similarities and Arithmetic. It also has a higher validity than any of the four verbal subtests of the Army Wechsler, excepting Arithmetic.

The Wechsler-Bellevue appears to have a somewhat better validity (.579, total scale) than the Army Wechsler (.553, total scale). This difference may, of course, be spurious. The inclusion of the relatively valid Digit Span test in the original version would tend to maximize the validity of the total scale used, especially if the intercorrelations of this test with the others were not high. It is probable that the validities of the two scales are approximately the same, when validity is defined as association with the criterion of the trainee's disposition in an Army Special Training Center.

One significant inference may, perhaps, be drawn from the data herein presented. It is that a quite valid scale which has been standardized upon the total range of intellect in a civilian population is also valid for the restricted mentality found among Army illiterates. It appears that revising such a scale for military use does not necessarily increase the validating coefficients to an appreciable degree.

Received January 2, 1945.

Use of the Shipley-Hartford Test in Evaluating Intellectual Functioning of Neuropsychiatric Patients *

M. Erik Wright, Lt. (jg) H(S), USNR

U. S. Naval Hospital, Oakland, California

The Shipley-Hartford Retreat Test ¹ was designed as an aid in detecting mild degrees of intellectual impairment in individuals of dull normal or higher original intelligence. The test also yields an estimate of the present level of intellectual functioning as well as an inference as to the prior level. The questions of intellectual level and of impairment are often significant in both the diagnostic and prognostic phases of neuropsychiatric case-work.

This study has two purposes: (1) To survey some of the intellectual abilities of a sample of hospitalized service personnel with neuropsychiatric involvements, and (2) to determine the validity of the Shipley-Hartford Test as a basis for estimating intellectual level.

Subjects and Procedure

The subjects were 977 patients ² admitted to the Neuropsychiatric Service of a mainland Naval Hospital who had been routinely examined with the Shipley-Hartford Test during the first fortnight after admission. Most of these men had seen overseas duty, with a large proportion only recently returned from active combat areas. Special referrals for a more extensive intelligence examination (Wechsler-Bellevue Test) were made for 134 of these patients within a few weeks of the first test. In Table 1 we have presented the age distribution of the 977 subjects. The average age was 27 years, and the range from 17 years to 64 years.

The educational achievement of the group is shown in Table 2. Since original school records were not available the patient's own report was used. Although such unconfirmed statements are subject to error,

* This article has been released for publication by the Division of Publications of the Bureau of Medicine and Surgery of the United States Navy. The opinions and views set forth in this article are those of the writer and are not to be considered as reflecting the policies of the Navy Department.

¹ Shipley, W. C. A self administering scale for measuring intellectual impairment and deterioration. *J. Psychol.*, 1940, 9, 371-377.

² The distribution of the group according to neuropsychiatric classifications cannot be published at this time because of war considerations.

Table 1
Age Distribution of a Randomly Selected Group of Neuropsychiatric Patients
Admitted to a Naval Hospital. N = 977

Age	17-19	20-24	25-29	30-34	35-39	40+
Per Cent of Patients	16	38	17	10	9	9

they permit a general estimate of the educational background of the subjects.

The range in education achievement was from second grade to the master's degree, with the average at the 10th grade. Almost 70% of the total group went beyond elementary school and a third either completed high school or went on to college. This is somewhat superior achievement to that characteristic of the population as a whole and may be due

Table 2
Educational Achievement (own report) of a Randomly Selected Population of
Neuropsychiatric Patients. N = 977

Highest Grade Completed	2-7	8	9-11	12	13-15	16+
Per Cent of Group	15	16	38	23	7	2

to the greater educational opportunities of the younger groups, to a tendency to over-estimate their achievement, etc.

Derivation of Scores

The Shipley-Hartford Retreat scale consists of two parts, a vocabulary test and an abstractions test. The vocabulary section is set up as a multiple choice test in which one of four alternatives has to be matched with the key word for best similarity. The abstraction test consists of unfinished problems. The subject is required to abstract the principle necessary to complete each one of them (e.g. 1 2, 3 5, 5 8 (7 1)).

Four scores may be obtained from the Shipley-Hartford, the vocabulary score, the abstractions score, the total score and the conceptual quotient (CQ). Age norms have been determined for the first three of these. The total score is the sum of the vocabulary and abstraction scores.

Conceptual Quotient is defined by the test constructor as follows:³ "The CQ (conceptual quotient) Scale is based on the clinico-experimental

³ Manual of Directions. *Shipley-Hartford Retreat Scale*. Published by the Neuropsychiatric Institute of the Hartford Retreat, Hartford, Conn., 1940 (p. 2).

observation that in mild degrees of mental deterioration, and in other conditions involving intellectual impairment, vocabulary is relatively unaffected, but the capacity for abstract (conceptual) thinking declines rapidly. . . . Impairment is measured by the extent to which the individual's abstract thinking falls short of his vocabulary. This deficit is expressed conveniently in the CQ (conceptual quotient)."

A table is presented in the Manual whereby the CQ may be obtained from the given vocabulary and abstractions scores.

"Original Intelligence"

Performance on the vocabulary test may be used as an approximation of "original" intelligence ⁴ for two reasons. First the vocabulary ability

Table 3

Performance on the Vocabulary Section of the Shipley-Hartford Test by
977 Patients of a Neuropsychiatric Service

Raw Score	Per Cent of all Patients	Age Equivalents	Intellectual Level*
37-40	3.1	19.8-21.0	Very Superior
33-36	8.2	18.2-19.4	Superior
29-32	20.8	16.6-17.8	High Average
25-28	22.8	15.1-16.2	Average
21-24	19.8	13.5-14.7	Average
17-20	12.2	11.9-13.1	Low Average
13-16	8.2	10.3-11.5	Borderline
11-12	2.7	9.5- 9.9	Mental Deficiency
1-10	2.4	Below 9.5	Mental Deficiency

* The age equivalents may be considered as mental ages. These have been translated in terms of intellectual level on the basis of Terman and Merrill's classification and distribution of adult mental ages.

seems more resistant to change than the ability to do abstractions. Secondly, many studies have shown that the performance on vocabulary correlates more highly with general tests of intelligence than does performance of any other single test. In Table 3 the vocabulary scores and their age equivalents for the 977 patients included in this study are presented.

The mean vocabulary score was 24.8 which is equivalent to a vocabulary age of 15. Two-thirds of the patients (68%) fall within one sigma of the mean (Vocabulary Score from 17.9 to 31.7) and 97% fall within two

⁴ The term, original intelligence, as here used, refers to the intellectual level prior to impairment and has no implications for the nurture problem.

standard deviations (Vocabulary Score from 11.0 to 38.6). This indicates that the distribution closely approximates a normal distribution curve.

Table 3 also shows that three-fourths (74.5%) of this group of patients were of average or above average intelligence. This compares very favorably with Wechsler's estimate of 75% based on a random sampling of approximately 1,000 adults in which a much more elaborate and refined testing instrument was used. From this it may be inferred that the basic intelligence of men who later became neuropsychiatric casualties is essentially the same as that of a random sampling of the adult population. However, a random sampling of individuals in the naval service who are not neuropsychiatric patients might reveal significant differences. Unfortunately, such comparative data are not as yet available.

Functional Intelligence

In order to distinguish "original intelligence" from the present intellectual level of an individual, the term "functional intelligence" is introduced. The usefulness of the Shipley-Hartford Test as a measure of functional intelligence may be determined by correlating its results with those on a test whose validity has already been established. For this purpose, a sample of 134 of the total group studied were administered both the Shipley-Hartford and the Wechsler-Bellevue tests. The total scores on the Shipley-Hartford (abstraction plus vocabulary) were correlated with the total scores on the Wechsler-Bellevue.⁵ The resulting correlation, $r = .77 \pm .03$ is as high as most of the correlations between two individual tests of intelligence, and is particularly good in light of the many differences between a group and individual test of intelligence (as time, range of abilities tested, administration, etc.). Thus, the use of the Shipley-Hartford test as a rough, but easily determined, approximation of the general intellectual level of the individual, when conditions do not permit the use of the more refined individual intelligence examination, is supported by these data.

The results on the distribution of functional intelligence as presented in Table 4 show that many of the patients are functioning considerably below their original intelligence. Whereas, the mean original intelligence score is very close to the upper limit of the average range, the mean functional intelligence score of 43 is very close to the lower limit of the average range.

⁵ The correlation of $r = .64 \pm .03$ between the Wechsler-Bellevue Verbal Score and the Shipley-Hartford Vocabulary score showed considerable overlap between the two tests, but the verbal and vocabulary scores alone seem less adequate as a basis for estimating the general level of intelligence.

The vocabulary based estimate showed that approximately 75% of the total group were either of average or above average intelligence. In contrast to this, the total score estimate indicates that only 53% of the group are functioning at that level. The difference is most striking in the lower intelligence levels. On the basis of "original" intelligence, 13%

Table 4

Functional Intelligence of a Group of Neuropsychiatric Patients — Measured
by the Total Scores on the Shipley-Hartford Test

Raw Score	%	Age Equivalents	IQ Description
74-80	.7	19.6-20.8	Very Superior
66-73	6.8	18.0-19.5	Superior
57-65	16.0	16.5-17.9	High Average
50-56	14.0	15.1-16.3	Average
42-49	15.5	13.5-14.9	Average
35-41	15.4	12.1-13.3	Low Average
25-34	20.2	10.4-11.9	Borderline
24 and Below	11.7	10.2 and Below	Mental Deficiency

of the group were characterized as being of borderline intelligence or mentally deficient. More than twice this number, 32%, are "functioning" at these levels.

Intellectual Efficiency

The concept of intellectual impairment suggested by the above discussion of original and functional intelligence has been treated by Shipley in terms of a conceptual quotient (CQ) which is the relationship of the abstract to the vocabulary abilities. The theoretical basis for the CQ is supported by a number of investigations.

The ability to formulate a principle which "abstracts" common elements from a group of events has been found to be a more complex psychological process than the type of discrimination between two objects which is so large a component of vocabulary skill. The ability to think in abstract terms or learn abstract relationships has also been shown to be a later development in the life of the individual ^{6,7} and is more readily

⁶ Straus, A. A., and Werner, H. Disorders of conceptual thinking in the brain-injured child. *J. Nerv. Ment. Dis.*, 1942, 96, 153-172.

⁷ Babcock, H. An experiment in the measurement of intellectual deterioration. *Arch. Psychol.*, 1930, No. 117, 1-105.

disturbed by psychopathological⁸ and organic brain disorders^{9, 10, 11, 12} than the more concrete abilities.

As the CQ decreases, the likelihood of intellectual impairment becomes more probable. The results show that a substantial proportion (62%) of the neuropsychiatric group had CQ's suggesting intellectual impairment. Using the interpretation of CQ's offered by Shipley, 10% of the patients had "slightly suspicious," 14% "moderately suspicious," 26% "very suspicious," and 12% "probably pathological" evidence of intellectual impairment. This is consistent with the clinical observations that neuropsychiatric patients are frequently unable to utilize adequately their intellectual capabilities to make plans, decisions, or deal with their problems in general.

Summary and Conclusions

The Shipley-Hartford Scale for the Measurement of Intellectual Impairment was given to 977 randomly selected Neuropsychiatric patients at a service hospital. Of this group, 134 were also given the Wechsler-Bellevue Intelligence test. The following conclusions seem warranted:

1. The Shipley-Hartford test can be used as a rough estimate of functional intelligence.

2. The distribution curve of "original" intelligence of neuropsychiatric patients is very similar to that obtained by Wechsler in a sampling of the adult population.

3. A large proportion (62%) of neuropsychiatric patients tend to show a lowering of efficiency in their intellectual functioning.

Received January 12, 1945.

⁸ Kendig, I., and Richmond, W. V. *Psychological studies in Dementia Praecox*. Ann Arbor: Edwards Bros., Inc., 1940, 1-166.

⁹ Goldstein, K., and Sheerer, M. Abstract and concrete behavior; an experimental study with special tests. *Psychol. Monogr.*, 1941, 53, No. 2 (Whole No. 239).

¹⁰ Hunt, H. F. A practical clinical test for organic brain damage. *J. appl. Psychol.*, 1943, 27, 375-386.

¹¹ Hunt, H. F. A note on the clinical use of the Hunt-Minnesota test for organic brain damage. *J. appl. Psychol.*, 1944, 28, 175-178.

¹² Hunt, H. F. *The Hunt-Minnesota test for organic brain damage*. Minneapolis: The University of Minnesota Press, 1943.

A Social I.E. Scale for the Minnesota Multiphasic Personality Inventory

Lewis E. Drake

University of Wisconsin

The Minnesota T-S-E Inventory (1) has been used as a part of the standard battery of tests administered to students in the guidance program at the University of Wisconsin for over a year. The inventory has yielded data which has been helpful in counseling students. Since, however, the Minnesota Multiphasic Personality Inventory (2) is also a part of the standard battery and since many items of the latter resemble items in the former, it was thought desirable to try to devise keys to score the Multiphasic to yield data now obtained by means of the T-S-E Inventory. This report is limited to results obtained for a Social I.E. scale. Scales for Thinking and Emotional introversion-extroversion are not yet ready for publication.

Procedure

An Item Analysis of the Multiphasic Personality Inventory was made by contrasting the percentage responses of two groups of students to the items. One group consisted of 50 students who obtained centile ranks of 65 and above on the T-S-E Inventory when scored for Social introversion-extroversion. The second group consisted of 50 students who obtained centile ranks below 35 on the T-S-E Inventory. The students were all females because of the small male population in the University, but the scale was validated with a male population as will be shown later. There was no other factor used in the selection of cases except that three cases were not included because the L scores on the Multiphasic were quite high.

Items were selected for the key which showed a difference between the percentage responses of the upper and lower groups of at least twice the standard error of the difference. Some significant items, however, were eliminated because there was an extremely high or extremely low frequency of response for both upper and lower groups.

After the item selection had been completed a new group of Multiphasic record sheets were scored with the obtained key for purposes of validation. These record sheets contained the responses of a group of female students who cleared through the testing office after the group of

students who provided the data for the item analyses. The scores obtained with the new key were then correlated with the Social I.E. scores obtained on the T-S-E Inventory. The key was then used for scoring all available record sheets for male students, providing there were T-S-E scores also available, and these scores were correlated with the T-S-E scores.

Finally, norms were established by scoring all Multiphasic record sheets available. The norms are reported in terms of T scores obtained in the customary way, namely: $T = 50 + \frac{10 (X_i - \bar{X})}{S}$ where X_i is the raw score, and \bar{X} the mean and S the standard deviation of the raw scores for the normative group.

Results

The items for this key are listed in Table 1 according to the way they are designated on the Multiphasic Record Sheets. The raw score is ob-

Table 1
Scoring Key for Social I.E.

A36	X	D50	X	E36	X	F8	X	H18	0
A37	X	D53	X	E38	X	F9	X	H51	X
A38	X	D54	X	E43	X	F30	0	H52	X
B6	X	E18	X	E44	X	F31	X	I21	0
B22	X	E23	X	E46	X	F34	X	I25	X
C20	X	E26	X	E47	X	F36	X	I26	X
C25	X	E27	X	E49	X	F41	X	I27	X
C48	0	E28	X	E52	X	F45	X	I28	X
C55	X	E29	X	E55	X	G18	X	I29	X
D2	0	E30	X	F2	X	G24	0	I38	X
D34	0	E32	X	F3	X	G35	X	I41	X
D35	0	E33	X	F4	X	G42	0	J24	0
D37	0	E34	X	F5	X	H2	X	J32	0
D45	0	E35	X	F6	X	H12	0	J33	X

tained by counting one point for every cell on the record sheet having an X corresponding to the key and one point for every cell which is blank corresponding to a 0 on the key. The cells containing question marks are not counted (3).

Twenty-eight of the items on this key have not been used on any keys reported by Hathaway and McKinley.

Record sheets for 87 female students were then scored with this key and the scores were correlated with the Social I.E. scores on the T-S-E. The resulting coefficient of correlation was $-.72$. The coefficient was negative because the key for the Multiphasic was constructed so that

high score would indicate introversion whereas on the T-S-E a low score indicates introversion.

Record sheets for 81 men students were likewise scored and the scores correlated with Social I.E. scores on the T-S-E. The resulting coefficient was $-.71$. Hence the key was used for both male and female students in obtaining norms.

Table 2 gives the T scores for this scale based upon records for 350 female students and 193 male students. Separate norms were computed

Table 2
The T Scores for the Social I.E. Scale

Raw Score	T Score	Raw Score	T Score	Raw Score	T Score	Raw Score	T Score
70	97	52	79	34	61	16	41
69	96	51	78	33	60	15	40
68	95	50	77	32	58	14	39
67	94	49	76	31	56	13	38
66	93	48	75	30	55	12	37
65	92	47	74	29	54	11	36
64	91	46	73	28	53	10	35
63	90	45	72	27	52	9	34
62	89	44	71	26	51	8	33
61	88	43	70	25	50	7	32
60	87	42	69	24	49	6	30
59	86	41	68	23	48	5	29
58	85	40	67	22	47	4	28
57	84	39	66	21	46	3	27
56	83	38	65	20	45	2	26
55	82	37	64	19	44	1	25
54	81	36	63	18	43		
53	80	35	62	17	42		

for males and females, but they were so similar, differing by only 2 from raw score 0 to 6 and being identical for most of the range, that the tables were combined for both sexes.

Summary

1. Using the Social I.E. scores on the Minnesota T-S-E Inventory for a group of female students as a criterion, an item analysis of the Minnesota Multiphasic Personality Inventory was made.

2. The derived key appears to have equally good validity for both male and female students.

3. An attempt is being made to derive Thinking and Emotional I.E. scales in a similar way.

Received July 5, 1945.

References

1. Evans, Catharine, and McConnell, T. R. A new measure of introversion-extroversion. *J. Psychol.*, 1941, 12, 111-124.
2. Hathaway, S. R., and McKinley, J. C. A Multiphasic Personality Schedule (Minnesota): I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249-254.
3. McKinley, J. C., and Hathaway, S. R. A Multiphasic Personality Schedule (Minnesota): II. A differential study of hypochondriasis. *J. Psychol.*, 1940, 10, 255-268.

Interests of Senior and Junior Public Administrators

Edward K. Strong, Jr.

Stanford University, California

Is it possible to differentiate junior and senior public administrators on the basis of their interests? By junior and senior administrators we mean roughly those earning three to five thousand dollars a year and those earning nine thousand and over.

The Committee on Public Administration of the Social Science Research Council obtained for the writer 550 Vocational Interest Blanks filled out by public administrators who were, in the judgment of the Committee, successful administrators. Some of these blanks for one reason or another could not be used. A few additional blanks were supplied from our files. For certain comparisons use has been made of blanks of city school superintendents but these blanks have not been included in general summaries respecting public administrators.

The public administrators have been classified in two different ways. First, they were classified into sub-groups according to the function they perform, such as, welfare, personnel, taxation, etc. Some of the administrators could not readily be classified in this manner for the reason that they direct employees engaged in many different activities, as for example, hospital superintendent, city manager and senior official in the Department of Agriculture. Under this first classification we have (a) functional and (b) general manager sub-groups as listed in Table 1.

Managerial Responsibility

The second classification, with which we are primarily concerned in this article, relates to the degree of managerial responsibility exercised by the administrator. The 550 cases ¹ were assigned as far as possible to five classes representing such degrees of managerial responsibility. As there were only a few men assigned to the lowest class it was discarded.

In determining the degree of managerial responsibility the following factors were taken into account: i.e., (1) Number of employees, (2) Whether employees were all engaged in the same type of work or in a

¹ As shown in Table 1 only 518 cases were used. Of the 550 cases, 10 were obviously not administrators, one turned out to be a duplicate, 10 evidently did not earn \$3,000, and 17 could not be classified. Six cases were added from our files, making actually 518 cases used from among 556.

Table 1

Classification of Public Administrators according to (1) Function Performed
and (2) Managerial Rank

	Managerial Rank				Total
	A	B	C	D	
Functional Groups					
Personnel	3	2	29	25	59
Social Insurance	3	1	4	3	11
Welfare	5	9	26	8	48
Taxation	0	0	6	9	15
Comptroller-Finance	1	4	13	6	24
Recreation and Parks	0	1	7	5	13
Office Manager	0	1	2	2	5
Statistics	1	3	21	3	28
Public Health	0	8	11	6	25
Engineer	1	9	14	2	26
Chemist-Physicist	0	3	11	0	14
Miscellaneous ^a	4	8	18	10	40
General Manager Groups					
Prison Warden	0	0	16	0	16
Hospital Superintendent	6	17	12	10	45
Reform School Superintendent	0	0	0	9	9
City Manager	1	5	8	27	41
Dept. of Agriculture and Commerce and TVA	9	30	13	1	53
Forest Service	0	11	20	15	46
Total	34	112	231	141	518

^a Includes 8 Publicity, 6 Law enforcement, 6 Education, 5 Lawyer, and 15 others.

variety of activities, (3) Whether employees were engaged in relatively simple or highly technical work, (4) Whether the position was essentially a line or a staff position, (5) Whether the administrator was in charge of the unit or was (a) an assistant or deputy administrator or was (b) an assistant to the administrator.

After the majority of public administrators had been assigned to one of the four classes it was found that salaries of these men approximated the following: Class A \$9,000 and up; Class B 7,000 to 8,999; Class C 5,000 to 6,999; and Class D 3,000 to 4,999. Amount of salary was then used as an additional factor in classifying the men. Some of those already assigned a class but whose salary was out of line proved to have been classified on really insufficient evidence and were reclassified on the basis of salary. The same basis was used in cases which had not been classified at first, because there was insufficient information concerning their work.

After making the functional classification and before making this managerial responsibility classification the writer corresponded with many of the men whose records were incomplete regarding the nature of their work. In most cases the desired information was obtained from them or from certain officials. In some cases salaries were secured when no other information was forthcoming. The additional information indicated that only a few had been incorrectly classified in terms of function and added considerably to our understanding of the relative importance of the duties of these men.]

Positions in some organizations are notoriously underpaid in comparison to other organizations; some younger men carry heavy responsibilities without commensurate salary and some older men have relinquished managerial responsibility for advisory work without decrease in pay. An attempt has been made to take such complications into account.

For convenience the classification is given in terms of salary but it must be recognized that the four classes represent degrees of responsibility rather than actual amounts of money received, although in most cases the man actually receives the salary of the class to which he is assigned.

The writer found the most difficult groups to classify were superintendents of hospitals and of reform schools, and prison wardens. Their classification is pretty much a guess. The writer is certain now that he was too much influenced by salary received by hospital superintendents. None of them it seems should have been assigned to Class A considering the calibre and scope of work performed by others in this class.

It is quite likely that some men have been assigned to the class above or below that to which they properly belong but we doubt if any man has been assigned to a class two steps above or below the class to which he belongs. Considering the complexities of the task and the lack of detailed information in many cases, the writer believes his classification is good. It is, in his opinion, much better than he believed possible at the beginning of the study.

Table 1 gives the number of public administrators from each functional group that was assigned to Classes A to D. Since the four classes are composed of different proportions of these functional groups and since such functional groups differ appreciably in interests it was feared that the summaries based on these four classes would be unduly influenced by the uneven representation. Mean scores for each class were calculated in the usual way and also by weighting the groups proportionately. Approximately the same results were obtained by both methods. We have used, however, the weighted means, except in the calculation of critical ratios of differences between means.

Table 2

I. Mean Interest Scores of Senior and Junior Administrators (Classes A and D) and Presidents of Manufacturing Concerns.
 II. Differences in Scores Between Classes A and D and Presidents. III. Mean Occupational Level Interest (OL)
 Scores for Occupations to be Compared with Differences in Scores Between Classes A and D

Group	Occupational Scale	Mean Scores of			Differences Between Scores of			Mean OL Scores of Occupations
		Class A	Class D	Pres.	A and D	D and Pres.	A and Pres.	
XI	President	38.2	31.2	49.6	7.0	18.4	-11.4	63.4
X	Journalist	36.5	30.0	29.4	6.5	-	7.1	63.0
	Advertiser	36.7	30.4	33.3	6.3	2.9	3.4	63.8
	Lawyer	42.2	33.7	29.8	8.5	-	12.4	64.4
IX	Realtor	32.3	29.4	37.9	2.9	8.5	-	60.4
	Sales Mgr.	33.3	30.1	39.2	3.2	9.1	-	63.3
	Life Ins.	29.1	27.0	33.0	2.1	6.0	-	62.3
VII	C. P. A.	33.9	31.2	26.5	2.7	-	7.4	63.4
	Artist	26.2	20.7	21.5	5.5	.8	4.7	58.9
	Architect	29.3	24.0	24.7	5.3	.7	4.6	61.0
	Psychologist	27.3	24.6	10.7	2.7	-13.9	16.6	60.9
	Physician	30.1	28.2	24.5	1.9	-	5.6	61.3
	Dentist	19.3	22.6	22.3	-3.3	-	-	57.7
II	Mathematician	28.6	24.4	17.9	4.2	-	10.7	61.5
	Engineer	33.6	32.4	32.9	1.2	.5	.7	61.4
	Chemist	30.3	30.4	23.3	-0.1	-	7.0	60.0
VI	Musician	22.0	23.1	14.4	-1.1	-	7.6	53.8

Table 2—Continued

Group	Occupational Scale	Mean Scores of			Differences Between Scores of			Mean OL Scores of Occupations
		Class A	Class D	Pres.	A and D	D and Pres.	A and Pres.	
V	City Sch. Supt.	36.0	34.9	19.1	1.1	-15.8	16.9	63.4
	Personnel Mgr.	39.2	39.9	27.6	-0.7	-12.3	11.6	61.4
	Minister	24.3	25.9	13.0	-1.6	-12.9	11.3	58.8
	Soc. Sci. Tchr.	29.0	33.6	19.6	-4.6	-14.0	9.4	56.1
III	Y.M.C.A. Sec'y	25.0	30.0	20.0	-5.0	-10.0	5.0	59.4
	Y Phys. Direct.	21.8	29.8	17.7	-8.0	-12.1	4.1	55.8
	Production Mgr.	34.2	38.9	40.3	-4.7	1.4	-6.1	60.2
	Purch. Agent	27.8	29.9	38.0	-2.1	8.1	-10.2	60.0
VIII	Banker	27.7	31.6	31.4	-3.9	-0.2	-3.7	58.1
	Accountant	27.0	34.4	30.4	-7.4	-4.0	-3.4	59.5
	Office Man	24.7	32.9	33.4	-8.2	.5	-8.7	57.0
	Aviator	22.1	28.6	—	-6.5	—	—	54.3
IV	Farmer	26.4	33.1	28.9	-6.7	-4.2	-2.5	55.7
	Math.-Sci. Tchr.	27.9	34.9	23.4	-7.0	-11.5	4.5	55.0
	Printer	22.9	30.1	22.6	-7.2	-7.5	.3	51.5
	Carpenter	10.1	19.8	16.7	-9.7	-3.1	-6.6	48.5
	Policeman	21.4	31.8	22.7	-10.4	-9.1	-1.3	50.0
	Forest Serv.	22.5	33.3	18.8	-10.8	-14.5	3.7	56.4
	OL	64.9	58.1	63.4	6.8	5.3	1.5	—
	MF	42.8	47.4	51.7	-4.6	4.3	-8.9	—

Interests of Senior and Junior Administrators

The mean scores on 34 occupational interests are given in Table 2 for Classes A and D. The data for Classes B and C are not published for their scores fall between those of A and D in 52 out of 70 cases. Such a relationship is illustrated by the artist interest scale, where Class A scores 26.2 on this scale; Class B scores 23.2; Class C, 22.2; and Class D, 20.7. In 17 of the remaining 18 cases the scores of Groups B and C deviate from those of A or D by less than 1.6 score. The greatest exception is in the case of city school superintendent interest, where Class A scores 36.0 on this scale; Class B scores 33.7; Class C, 32.3; and Class D, 34.9. This is the only case among 70 where the score of either Class B or C falls outside the scores of Groups A and D by an amount approximating a statistically significant difference (critical ratio of 2.4). Consequently we are justified in assuming that as far as interests go the four groups constitute a continuum and that differences between Groups A and D reflect differences in the four groups. The intercorrelations between the interest profiles of the four classes, as given in Table 4, further support this statement. (Such an array of data affords good evidence that the classification into the four groups has real merit.)

The fact that the scores in Class B fall so uniformly between Classes

Table 3
High Interest Scores of Four Classes of Administrators and of Presidents of Manufacturing Concerns

Group	A Ratings 45 and up	B+ Ratings 40 to 44	B Ratings 35 to 39
Class A		Lawyer	Personnel manager President Advertiser Journalist City School Supt.
Class B		Personnel manager	Production manager Lawyer
Class C			Personnel manager Production manager Lawyer
Class D		Personnel manager	Production manager Math.-Science teacher City School Supt.
President	President	Production mgr.	Sales manager Realtor Purchasing agent

A and D and that they differ very little from Class A justifies us in basing conclusions on Class A to a greater extent than is warranted by the small number of cases included in it, i.e., 34.

When only the high interest scores in Table 2 are considered we have them as shown in Table 3. We have included here the corresponding data for presidents of manufacturing concerns for comparison. The data in the table concerning Classes A and D and presidents are not too easily appreciated but as soon as these scores are shown on the interest globe (Fig. 1) their significance becomes apparent. The strongest interests of senior administrators are in president of Group XI, advertiser, lawyer and journalist of Group X and personnel manager and city school administrator of Group V, all located near the top of the right-hand figure. The strongest interests of Class D are in production manager of Group III, math.-science teacher of Group IV and personnel manager and city school superintendent of Group V, located across the bottom of the same figure and extending upwards at the right to overlap with the interests of Class A in Group V. The strongest interests of presidents are at the left hand side of the figure including realtor and sales manager of Group IX, purchasing agent of Group VIII, president in which they overlap with Class A, and production manager in which they overlap with Class D. Each of the three have some strong interests in common with the other two and each has some interests peculiar to itself.²

The rank-order correlations between the interest profiles of Classes A and D and president are: Class A and Class D = .43; Class A and president = .59; and Class D and president = .39. These are low coefficients as this type of correlation goes (see Table 4). They are about equal to the correlations between the personnel and publicity functional groups and between city school superintendents and forest service administrators, whose interests differ appreciably. As far as the three coefficients go they suggest that senior administrators and presidents are more similar in their interests than either is similar to junior administrators.

If we include in our comparison not only the high ratings of A, B+ and B but also the B- ratings we can add that senior administrators have more the interests of presidents, of men engaged in selling and influencing people and of scientists, whereas junior administrators have more the interests of social workers, production managers, general office people and skilled workmen. On the same basis we can say that senior administrators differ from presidents by having more of the interests of men engaged in social work and in influencing people but not in selling them; by having

² The strongest interests of both Classes B and C are in personnel manager, production manager and lawyer having interests which fall between those of Classes A and D.

more the interests of scientists, particularly of psychologists and engineers; and of public accountants but not of office people, including accountants; and by having less of the interests of production managers and of presidents.

Several side lights regarding the four classes of administrators and how their interests are related to those of functional and general manager

Table 4

Rank-Order Correlations Between Interest Profiles of Classes A, B, C and D and of Certain Groups of Public Administrators and Business and Professional Men

	Classes			
	A	B	C	D
Class				
A	—	.84	.69	.43
B	.84	—	.92	.74
C	.69	.92	—	.87
D	.43	.74	.87	—
Functional Groups				
Personnel	.62	.84	.87	.76
Recreation	.21	.55	.58	.61
Statistics	.67	.63	.69	.69
Law	.41	.44	.32	.31
Chemist-Physicist	.35	.26	.32	.29
General Manager Groups				
Prison Warden	.56	.85	.88	.77
City Manager	.50	.72	.82	.74
District Ranger	-.21	.19	.31	.37
Forest Supervisor	.29	.71	.73	.64
Forest Serv. Administrator	.87	.87	.74	.55
Dept. of Agriculture Admin.	.86	.93	.85	.70
Dept. of Commerce Admin.	.61	.66	.74	.64
Business and Professional Men				
President	.59	.68	.67	.39
Production Manager	.37	.62	.71	.58
Personnel Manager	.61	.81	.83	.68
Engineer	.46	.41	.46	.27
Lawyer	.81	.77	.65	.31

groups are brought out in Table 4. The correlations in this table are all between interest profiles of different groups. The first four rows of coefficients show that Classes A, B, C and D stand in the order of a continuum, as previous data indicated, that Classes B and C are most closely related (coefficient of .92); second, C and D (.87); and third, A and B (.84); and that A and D are not closely related (.43).

Both the personnel functional group and personnel managers from

industry have interest patterns which agree particularly well with Classes B and C, although 91% of the personnel functional group are assigned to Classes C and D, and both these personnel groups correlate about .61 with Class A. On the other hand, the recreation group correlates much higher with Class D (.61) than with Class A (.21). The statistics, law and chemist-physicist functional groups correlate to about the same degree with all four classes, the coefficients being in the sixties for statisticians, and quite low for the other two groups.

Among the three groups of forest service men, the interests of district rangers are little related to any of the four classes, but more closely related to Class D (.37) than to Class A (-.21); the forest supervisors correlate highest with Class C (.73) and the administrators correlate highest with Classes A and B (.87). Relationships such as these are seemingly what should be expected and aid one in understanding what the other coefficients mean.

The interests of prison wardens and city managers are like those of forest supervisors in being more closely related to Class C than to the other three classes.

Eleven of the 34 men assigned to Class A are from the Department of Agriculture, 3 from the Department of Commerce and none from the Forest Service (which is a part of the Department of Agriculture but which has been kept as a separate group in this study). Nevertheless the Forest Service administrators appear to have interests more related to Class A than any other group (.87) with the Department of Agriculture practically tied with them (.86) and the Department of Commerce less closely related (.61).

Among business and professional men it is the lawyer who has interests most closely related to Class A (.81).³ Presidents' interests are more closely associated with Class B (.68) and production managers' and personnel managers' interests correlate highest with Class C. The engineers' interests are little related to any class just as is true of the functional group of chemists-physicists.

It seemed strange to the writer that the interests of Department of Commerce men should differ so much more from Classes A, B, C and D than the interests of men in the Department of Agriculture. Since the former include 24 men assigned to the statistics functional group and since these men have a rather peculiar assortment of interests it occurred to us that the rather low correlation might be caused by the presence of the statisticians. Accordingly correlations were calculated between the remainder of the Department of Commerce group, omitting the statisti-

³ All our data make clear that the five lawyers constituting our law functional group differ from lawyers in business.

cians, and Classes A, B, C and D. The latter four coefficients average .09 lower than the original coefficients given in Table 4. So it is not the presence of statisticians in this department that causes it to have interests less like Classes A, B, C and D than do the Department of Agriculture. The data in Table 5 indicate that the interests of the Department of Commerce administrators are more like presidents, production managers and engineers than the Department of Agriculture men and less like personnel managers and lawyers than the latter.

Table 5

Rank-Order Correlations Between the Interest Profiles of Business Men and of Administrators, from the Departments of Agriculture and Commerce

	Dept. of Agriculture	Dept. of Commerce
President	.47	.62
Production Manager	.38	.74
Personnel Manager	.63	.49
Engineer	.31	.87
Lawyer	.73	.42

Differentiation of Junior and Senior Administrators

The preceding data indicate that junior and senior administrators differ appreciably in their interest. It is not easy, however, to express the differences between the two by such complex relationships as are portrayed in Figure 1. Are there any short cuts that may be used for this purpose?

First of all let us note that such differentiation can not be obtained by use of the public administrator interest scale. The mean scores of the four classes of public administrators on the public administrator scale are: A, 52.1; B, 49.7; C, 49.3; and D, 48.8. Not even the difference in mean scores of Classes A and D is statistically significant (critical ratio of only 1.9). Differences in success-failure, or in this case, differences in senior-junior administrator standing, are not likely to be revealed by a scale based on both groups.⁴

From Table 2 we note that there are four scales on which senior administrators score significantly higher than junior administrators, that is, by scores of 6 or more. These scales are president, journalist, advertiser and lawyer. And there are ten scales on which senior administrators score lower than junior administrators. One way to decide

⁴ Scores of sub-groups of public administrators on the public administrator scale are given in: E. K. Strong, Jr., *Interests of public administrators*, *Public Personnel Review*, 1945, 6 166-173.

whether a man is a senior rather than a junior administrator it to note if he scores 35 or higher on the president, journalist, advertiser and lawyer scales and if he scores lower than 30 on the Y. M. C. A. physical director, accountant, office worker, aviator, farmer, math.-science teacher, printer, policeman and forest service scales, and lower than 20 on the carpenter scale.

Reference to Figure 1 will show the location of the four and ten occupational interests on the interest globe. The four comprise Groups X and XI near the top of the globe and the ten comprise Group IV and half of Group VIII which are near the bottom of the globe. Further reference to the figure and Table 2 will disclose that to a high degree as one goes from occupational interests at the top of the globe to those at the bottom one goes from large plus to large minus differences in score between senior and junior administrators. This relationship is expressed extremely well by scores on the OL (occupational interest level) scale.

Differentiation by OL Scale. The last column in Table 2 gives the mean OL score of the criterion groups upon which the 34 occupational scales are based. It will be noted that the four scales upon which senior administrators score significantly higher than junior administrators have OL scores ranging from 63.0 to 64.4 (average of 63.7) and that the ten scales on which senior administrators score significantly lower than junior administrators have OL scores ranging from 48.5 to 59.5 (average of 54.0). If we correlate all the differences in occupational level interest scores between senior and junior administrators (column four of Table 2) with the corresponding OL scores we obtain a rank coefficient of .84. Evidently the OL scale measures the interests which differentiate senior and junior administrators to a very considerable degree.

The OL interest scale contrasts the interests of business and professional men, typifying the upper socio-economic level, with the interests of common laborers, typifying the lower socio-economic level. Mean scores on this scale for 34 occupations are given in Table 2. The hierarchy of occupations based on interests is quite similar to the hierarchy based on intelligence test scores.⁵

The mean OL scores of the four classes of public administrators are shown in Table 6, with corresponding occupational groups.

The difference between the OL scores of Groups A and D, amounting to 6.0, does not appear large but it represents 30% of the entire range from the highest to the lowest socio-economic levels and has a critical ratio of 4.8. The data indicate that senior administrators score as high

⁵ E. K. Strong, Jr. *Vocational interests of men and women*. Stanford University Press, 1943, Chapter 10.

as any occupation so far tested and that junior administrators, as a group, belong to a lower socio-economic level.

If it could be shown that OL scores increase with age then it would be easy to explain the fact that Class A averages higher than Class D as Class A is composed of older men than Class D. Unfortunately for such a hypothesis the data so far accumulated indicate practically no increase in OL score with age.⁶ Table 7 gives the mean OL scores of Classes A to D for ages ranging between 25 and 70 years. There is here no evidence of increase in OL score with age.

Some of the men now in Classes C and D will eventually move into Classes B and A. But as far as the interests measured by the OL scale

Table 6
Mean OL Scores for Four Classes of Public Administrators

Public Administrators	Mean OL Score	Occupations with Comparable Score
Class A	64.9	{ Lawyer 64.4 President 63.4
Class B	61.9	{ Life insurance 62.3 Mathematician 61.5 Personnel manager 61.4
Class C	60.1	{ Production manager 60.2 Chemist 60.0 Purchasing agent 60.0
Class D	58.9	{ Accountant 59.5 Minister 58.8 Banker 58.1

go it is evident that a considerable number in Classes C and D cannot have scores as high as men in Classes A and B or there would not be the differences in mean scores which are given above. The actual distribution of OL scores for the four classes is given in Table 8. Assuming that senior administrators should not score below 55 on this scale we have: 5.9% of A administrators scoring below 55; 10.0% of B administrators scoring below 55; 24.7% of C administrators scoring below 55; and 25.0% of D administrators scoring below 55. In terms of overlapping between sub-group A and the other three sub-groups we have: 86.2% of B overlap with A; 73.8% of C overlap with A; and 67.1% of D overlap with A. On the basis of OL interest scores we can roughly estimate that between a fourth and a third of Group D do not have the interests characteristic of senior public administrators. Unfortunately we do not have similar comparisons in terms of ability to compare with these calculations in-

⁶ Ibid., Chapter 10.

volving interests. Such a statement does not imply that this minority are necessarily in the wrong type of work. Many Class D activities are sufficiently different from Class A activities to demand men of different interests and abilities for their successful handling.

Differentiation on the Lawyer Scale. Mean scores of the four classes of public administrators on the lawyer interest scale are as follows: A, 42.2; B, 36.2; C, 34.3; and D, 33.7. The critical ratio of the difference between senior and junior administrators is 4.4, slightly less than the critical ratio of 4.8 on the OL scale. The data suggest, however, that the lawyer scale differentiates Classes A and B better than the OL scale

Table 7

Distribution of OL Scores of Four Classes of Public Administrators According to Age

Age	Classes									
	A		B		C		D		Total	
	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
70					1	61			1	61
65	1	72	3	62	8	65	4	56	16	62
60	2	69	9	62	14	60	6	55	31	61
55	4	66	19	64	26	61	16	57	65	61
50	7	62	18	64	46	59	22	58	93	60
45	7	67	26	62	44	59	20	60	97	61
40	9	66	22	62	40	60	27	62	98	61
35	2	63	11	60	17	58	25	59	55	59
30	2	57	1	62	23	59	17	58	43	58
25			1	62	3	67	3	58	7	62
Total Cases	34		110		222		140		506	
Mean		65		62		60		59		60
Cases with no age given			2		9		1		12	

the corresponding critical ratios are 3.0 and 2.3 respectively, but the reverse is the situation between Classes C and D where the respective critical ratios are 0.6 and 1.5. Similar results are to be expected from the lawyer and OL scales since they correlate .60.

It might be supposed that this differentiation between the four classes in terms of lawyer interest is due to the fact that there are more legally trained men in Class A than Class D. The facts, however, do not justify this hypothesis. Of the 333 men who have supplied sufficient information about their education to determine what field they specialized in there are only 46 men reporting legal training. (We do not here count one or two courses in the subject but rather a preparation that would lead to legal practice if the man so desired.) This amounts to 13.8%.

The percentage is 12.3 for men in Classes A and B and 14.9 in Classes C and D. Consequently increase in lawyer interest scores from Class D to Class A is not attributable to increase in legal training from Class D to Class A.

Naturally the various functional sub-groups vary in mean lawyer score and in percentage of members with legal training. There is, however, some association between the two sets of data (correlation of .61). It is difficult to say whether a functional group has a high mean lawyer interest score because it has more than its share of men with legal training or that the type of work performed attracts men with interests of

Table 8

Distribution of OL Scores of Four Classes of Public Administrators.
Figures are Percentages

OL Score	A	B	C	D
80			0.4	
75	11.7	1.8	0.9	
70	14.7	10.9	7.7	8.6
65	17.6	20.9	20.3	13.6
60	35.3	35.4	23.9	24.3
55	14.7	20.9	22.1	28.6
50	5.9	9.1	16.2	17.9
45		0.9	6.3	4.3
40			1.8	1.4
35				0.7
30			0.4	0.7
Mean*	64.9	61.9	60.1	58.9
σ	6.9	6.1	7.7	7.4
N	34	112	231	141

* Critical ratios of differences in mean scores are as follows: AB, 2.3; AC, 3.8; AD, 4.8; BC, 2.4; BD, 3.5; and CD, 1.5.

lawyers and so men with legal training are more apt to be found in that work than in other activities.

Differentiation on the President Scale. Mean scores of the four classes of public administrators on the president scale are as follows: A, 38.2; B, 34.1; C, 33.3; and D, 31.2. In terms of critical ratios, senior and junior administrators are not differentiated as well as on the lawyer and OL scales, the three critical ratios are respectively 4.0, 4.4 and 4.8. Classes A and B are differentiated as well on the president scale as on the OL scale (C.R. of 2.3) but not as well as on the lawyer scale (C.R. of 3.0). The president scale, however, differentiates Classes C and D better than the other two scales, critical ratios, respectively, of 2.0, 0.6 and 1.5.

Summary and Conclusion

Over 500 public administrators have been classified on the basis of the managerial responsibility exercised by them. The four classes earn approximately \$3,000 to \$4,999; \$5,000 to \$6,999; \$7,000 to \$8,999 and \$9,000 and above. The first and fourth classes are referred to here as junior and senior public administrators.

The interests of senior and junior public administrators differ somewhat,—enough in fact to suggest that a fourth to a third of junior administrators do not have the interests of senior administrators. Such evidence as we have does not warrant the belief that the interests of the junior administrators will change with increasing age in the direction of senior administrators.

Senior public administrators and presidents of manufacturing concerns have interests more in common than has either with junior administrators. But senior administrators and presidents differ appreciably. The former have more than the latter of the interests of men engaged in social work and in influencing people but not in selling them, have more the interests of scientists, particularly of psychologists and engineers, and of public accountants and less of the interests of general office people including accountants, of production managers and of presidents.

Senior administrators differ from junior administrators by having more the interests of presidents, of men engaged in selling and influencing people and of scientists, whereas junior administrators have more of the interests of social workers, production managers, general office people and skilled workmen.

Senior and junior administrators differ significantly in their scores on fourteen occupational interest scales. It is possible that the two groups could be well differentiated by using some weighting system applied to scores on these scales. A general summary of such differences is measured by the occupational level (OL) scale, on which the two groups are differentiated by a critical ratio of 4.8. They are also differentiated significantly by the lawyer and president scales, which correlate with OL, by .60 and .63, respectively.

The two groups of administrators are not differentiated by scores on the public administrator scale which is based on their records and on those of the two intermediate classes of administrators. Such a scale measures the differences in interests between public administrators as a whole and men-in-general, representative of the upper socio-economic level. Degree of possession of managerial responsibility cannot be measured to any appreciable degree by such an interest scale.

Degree of success-failure is quite another measure from that of degree of possession of managerial responsibility. It cannot be determined

satisfactorily any more than the latter on an occupational interest scale. If success-failure is to be measured in terms of interests it must be accomplished by contrasting the interests of men who are successful with the interests of men who are not successful. Almost nothing has been done in this area so we do not know what are the possibilities.

Extensive data regarding men in the Forest Service indicate that as one goes upward from district ranger to the top administrative positions there is a progressive decrease in the interests typical of rangers and an equally progressive increase in the interests of administrators.⁷

Top administrators perform different work from that of lower officials and evidently possess somewhat different interests. Selection of men should be on a different basis for the top and bottom levels. As only a small percentage of men at the bottom ever reach the top it is necessary to select only a small percentage of men entering a profession who have the characteristics of men at the top. The remainder can be selected for the work they are to perform in the lower and middle levels. It seems obvious that there should be some other way of rewarding competent men who perform the middle level jobs successfully than by promoting them into administrative work for which neither their interests nor abilities fit them.

At the moment the OL scale seems to measure the differences in interests of junior and senior administrators as well as any scale. Further research is needed to substantiate this statement. Efforts should be made to see if some revision of the present OL scale may not perform this service even better. Comparison should be made between scores on an intelligence test and OL scores to see which are more useful here, or whether some combination of the two is better than either one alone in differentiating degrees of administrative ability.

Received January 29, 1945.

⁷ E. K. Strong, Jr. The interests of Forest Service men, *Educational and Psychological Measurement*, 1945, 5, 157-171.

A Comparison of the Thurstone and Likert Techniques of Attitude Scale Construction

Allen L. Edwards

University of Washington

and

Kathryn Claire Kenney

University of Maryland

This is a study in the methodology of attitude measurement; a comparison and evaluation of two methods of attitude scale construction. Although various techniques for the measurement of social attitudes have been suggested,¹ the two most frequently used methods are probably the "method of equal appearing intervals" developed by Thurstone and Chave (18) and the "method of summated ratings" ² developed by Likert (11). This study is concerned with the relative merits of the Thurstone and Likert techniques of scale construction.

Method of Equal Appearing Intervals

The method of equal appearing intervals begins with the collection of a variety of statements of opinion toward a particular issue which are then screened and edited in accordance with certain "informal criteria." Statements which appear to represent past rather than present attitudes are discarded or re-worded, as are statements which appear to be double-barreled ³ or which contain confusing or ambiguous concepts. Inspection should also exclude statements which might be approved by individuals with opposed attitudes.⁴

After the statements have been edited, they are then presented to a group of judges who are instructed to sort them into various categories to represent a scale ranging from extremely favorable, through neutral,

¹ Excellent summaries can be found in Albright (1, pp. 181-213), Bird (3, pp. 149-167), LaPiere and Farnsworth (10, pp. 397-399), and Murphy, Murphy, and Newcomb (13, pp. 891-912).

² The term "method of summated ratings" was introduced by Bird (3, p. 159) to describe the procedure followed by Likert.

³ But in this connection see the discussion by Edwards (4, p. 578) which points to the possible value of statements which contain a "rationalization" clause.

⁴ A more detailed enumeration of the rules to be followed can be found in the monograph by Thurstone and Chave (18, pp. 56-58).

to extremely unfavorable expressions of opinion about the issue or institution in question. The judges are not asked to give their own opinions, but merely to estimate the degree of favorableness or unfavorableness expressed by each statement. When the sorting procedure is completed, tabulations are made indicating the number of judges who placed each item in each category. From these data accumulative proportions are computed for each item and ogives are constructed. Scale values of the individual items are then read from the ogives, the value of each item being that point along the base line, in terms of scale value units, above and below which 50 per cent of the judges placed the item.

A statistical criterion of the ambiguity of the items is provided in terms of the width of the range between the points on the scale marking off the 25th and 75th percentiles. This distance is called the *Q* value. A small *Q* value indicates that the middle 50 per cent of the judgments spread over a relatively small range or, in other words, that there is a good deal of agreement among the judges as to where the item belongs on the scale. A large *Q* value indicates lack of agreement among the judges and, indirectly, that something is probably wrong with the wording of the statement.

Items are selected for the final scale on the basis of the computed scale and *Q* values. An attempt is made to select about 20 or 22 items with low *Q* values and with scale values falling at relatively equally-spaced distances along the continuum. Two comparable forms of the scale, in terms of scale and *Q* values of the items included in each form, are constructed. The two forms are then given to a new group of subjects who are asked to check those statements with which they agree. The score for the individual subject is the mean or median scale value of the items which he has checked as being those with which he agrees. Reliability of the scales is found by correlating scores on the two forms of the scale.

Method of Summated Ratings

The method of summated ratings also calls for a collection of various statements of opinion which are then edited in accordance with informal criteria similar to those used in the method of equal appearing intervals.⁵ After the elimination and editing of items failing to meet the prescribed standards, the remaining statements are presented to a group of subjects who are asked to respond to each one in terms of their own agreement or disagreement with the statement. Usually a 1 to 5 scale of response is used; subjects check whether they strongly agree, agree, are undecided,

⁵ See Murphy and Likert (12, pp. 281-283) and Rundquist and Sletto (16, pp. 6-8) for a discussion of these standards.

disagree, or strongly disagree with each statement. A score is given for each item depending upon the response made. The five possible responses may be weighted 1-2-3-4-5 or 5-4-3-2-1.⁶ Either 1 or 5 is consistently favorable or unfavorable, although the continuum is reversed in about half the statements. That is, about half the statements are worded so that a strongly agree response indicates a favorable reaction to the issue in question, while the other half of the statements are worded so that a strongly agree response indicates an unfavorable reaction. The score for the individual subject is the sum of all scores for the separate items.

In selecting items for the final scale, a criterion of internal consistency is used. Criterion groups consisting of the upper and lower 10 (or some other) per cent of the subjects in terms of total scores are compared to find whether the individual items will differentiate between the two groups. The means of the upper and lower groups for each item are found; items which show the largest difference between the means of the two groups are retained in the final scale.

Scales constructed by the method of summated ratings usually contain about 20 to 25 items, although Hall (8) has used scales with as few as 5, 7, and 10 items. Reliability of the scales is found by the split-half method of correlating scores for the odd versus even items.

Criticisms of Thurstone's Method

The monograph by Thurstone and Chave, describing in detail the method of equal appearing intervals for measuring attitudes, appeared in 1929. By the time Likert's monograph describing his technique appeared in 1932, the Thurstone procedure was generally recognized as a major, if not the most important, development in the field of attitude scale construction. It is important, therefore, if we are to compare the two methods, to examine the motivation behind Likert's departure from the by then already well-established Thurstone technique. Some indication of this is given by the following quotation from Murphy and Likert⁷:

"A number of statistical assumptions are made in the application of his (Thurstone's) attitude scales—e.g., that the scale values of the statements are independent of the attitude distribution of the readers who

⁶ This is a simplified method of scoring which was found to correlate .99 with the more complicated sigma method first used.

⁷ We quote from Murphy and Likert rather than from Likert's original report because the Murphy and Likert publication is probably more readily available to the interested reader and since it contains, with but few corrections or omissions, the material originally reported by Likert and, in addition, a more detailed report of applications of scales constructed by the technique. The passage quoted, with but minor changes, is the same as that appearing in Likert (11, p. 6).

sort the statements—assumptions which, as Thurstone points out, cannot always be verified. The method is, moreover, exceedingly laborious. It seems legitimate to inquire whether it actually does its work better than the simpler scales which may be employed, and in the same breath to ask also whether it is not possible to construct equally reliable scales without making unnecessary statistical assumptions” (12, p. 26).⁸

The main contentions of Murphy and Likert regarding the method of summated ratings seem essentially to be: (1) “it avoids the difficulties encountered when using a judging group to construct the scale” (12, p. 42); (2) “the construction of an attitude scale by the sigma method⁹ is much easier than by using a judging group to place the statements in piles from which the scale values must be calculated” (12, p. 43); (3) “it yields reliabilities as high as those obtained by other techniques with fewer items” (12, pp. 42–43); (4) it gives results which are comparable to those obtained by the Thurstone technique.¹⁰ More generally, the method of summated ratings “seems to avoid many of the shortcomings of existing methods of attitude measurement, but at the same time retains most of the advantages present in methods now used” (12, p. 42). These claims, it should be noted, have been vigorously contested, notably by Bird (3) and Ferguson (7). For our part, we shall, in the sections which follow, attempt to evaluate them in the light of available evidence.

Influence of the Judging Group

Several studies cast doubt upon Murphy and Likert’s criticism that the attitudes of the judging group may influence the scale values of items when the method of equal appearing intervals is used. Using various approaches to the problem, these studies (5), (9), (15), seem to be in agreement that the attitude of the judging group is not a seriously disturbing factor. Hineckley’s study (9), in particular, is clear cut. Groups of white students with differing attitudes toward the Negro were asked to sort items expressing opinions about the Negro. A high positive correlation was obtained between the scale values assigned to the items by the white students who were favorable and by those who were unfavorable in attitude toward the Negro. A high positive correlation was also obtained between scale values derived from judgments of an antagonistic white group and from the judgments of a group of Negroes.

⁸ The method of summated ratings also makes certain statistical assumptions as Murphy and Likert recognize (12, pp. 26 ff.). Cf. also the paper by Ferguson (6).

⁹ Later replaced by the even simpler 1 to 5 method.

¹⁰ We have failed to find a specific statement to this effect, yet the idea seems implied in the paragraph quoted above.

Simplicity of the Likert Method

Investigators who have used the Likert method seem to be in agreement that it is simpler than the method of equal appearing intervals. Hall reports that he used the method of summated ratings in his survey of the attitudes of employed and unemployed men "because of its relative simplicity and because it yields scales of high reliability" (8, p. 6), and Rundquist and Sletto, who used the Likert technique in constructing the Minnesota Survey of Opinions scales, agree that "it is less laborious than that developed by Thurstone" (16, p. 5). This evidence is, of course, in the nature of authoritative opinion, and Bird has raised some rather pertinent objections concerning it.

"Will the experimenter spend more time, too, in scoring every item and summing them in these long scales than another might spend determining the mean or median value by the Thurstone technique? Then too, is it actually less time-consuming to validate items in terms of selected groups than to determine the Q values from a curve or a distribution of scores? The claim of greater or lesser laboriousness seems to have been put forward without due regard for all processes in scaling techniques; but, in the interest of constructing refined measuring instruments, time can be neglected. There is much to be said in favor of a psychologist's refining his instrument before actually applying it to experimental groups. The argument that the method of summated ratings is less laborious limps badly" (3, p. 161).

Bird's points are well taken, particularly in the case of scales constructed by the method of summated ratings which contain more than, let us say, 25 items. But most investigators who have used the method of summated ratings have not found any need for "long scales" to which Bird is objecting. After our own experience in constructing *both* Likert and Thurstone scales, we are inclined to agree with other investigators that scales can be constructed by the method of summated ratings more quickly and with less labor than by the equal appearing interval method. We found, for example, that construction of the Thurstone scales required about twice as much time, exclusive of the time spent by the judging group in sorting the items, as did the Likert scales. It is unfortunate that this is but an estimate and that our records do not permit a more precise statement of the time factor—a point which should be checked in future research and reported.

Reliabilities of the Two Methods

A note of confusion has centered around the subject of reliability largely as a result of Likert's study of the reliability of a Thurstone-type scale which was scored by both his and the Thurstone technique. The

scale was given to a group of subjects with instructions to check the items in accordance with the usual Thurstone instructions. The same scale was then given to the subjects with instructions to check for each item one of the five alternatives (strongly agree, agree, undecided, disagree, strongly disagree) in accordance with the usual Likert instructions. Four of the items on the Thurstone scale were not adaptable to Likert-type responses and were omitted when the subjects were asked to check their reactions according to the method of summated ratings scoring system.

The reliability coefficient between the two forms of the scale (22 versus 22 items), when scored by the Thurstone method, was .88 (corrected). The reliability coefficient for the two forms (18 versus 18 items) as scored by the Likert method was .94 (corrected). What this demonstrates, of course, is that it is possible to take a scale constructed by the Thurstone technique and to apply to most of the items the Likert method of scoring. But one critic seems to think that because of this finding, Likert erroneously concluded that "his technique is the better one" (7, p. 52). The higher reliability coefficient obtained by the Likert method of scoring, he adds, may be due to the fact that "increasing the number of steps in a psychological scale increases reliability" (7, p. 52). As a matter of record, this is precisely the same explanation offered originally by Murphy and Likert (12, p. 55 and p. 47) for the higher reliability coefficient obtained by the 1 to 5 method of scoring. The entire discussion, pro and con on this point, it seems to us, has little bearing upon the question of method—the *method* of summated ratings or the *method* of equal appearing intervals will yield scales of higher reliability. The real problem concerns the reliabilities of scales constructed by the two methods, not the reliability of a particular scoring scheme isolated from the technique of scale construction of which it is a part. And on this question there is ample evidence.

Ferguson has quoted Thurstone as reporting the reliabilities of scales constructed by the method of equal appearing intervals, under his editorship, as being "all over .8, most of them being over .9" (6, p. 670). We do not know whether these coefficients are for scales of 20 or 40 items, but Ferguson mentions that in his own studies he has found reliabilities for Thurstone scales ranging from ".52 to .80 for the 20-item forms and from .68 to .89 for the 40-item forms" (6, p. 670). If we take these coefficients as representative, how do they compare with those reported for scales constructed by the method of summated ratings?

Murphy and Likert found reliability coefficients for their Internationalism Scale of 24 items ranging from .81 to .90.¹¹ Their Imperialism

¹¹ The reliability coefficients of the Likert scales are based upon split-half correlations, and all of those reported here have been "corrected" to indicate the reliability

Scale of 12 items gave coefficients ranging from .80 to .92; the Negro Scale of 14 items yielded coefficients ranging from .79 to .91 (12, p. 48). Rundquist and Sletto report coefficients ranging from .78 to .88 for various scales of 22 items each (16, p. 110).

That Likert-type scales with even fewer items will give high reliability coefficients is indicated by Hall. Reliability coefficients for his religious scale of 10 items ranged from .91 to .93; for the scale of 7 items measuring attitude toward employers the coefficients ranged from .77 to .87; and the morale scale of 5 items gave coefficients of .69 to .84 (8, p. 19). All of these coefficients compare favorably with those obtained from scales constructed by the method of equal appearing intervals. According to the evidence at hand, there is no longer any reason to doubt that scales constructed by the method of summated ratings and containing fewer items will yield reliability coefficients as high as or higher than those obtained with scales constructed by the Thurstone method.

The Need for a Judging Group

The confusion which followed Likert's re-scoring of a Thurstone-type scale by the 1 to 5 method, unfortunately, has not been confined to the subject of reliability; it has spread to involve the question of whether or not there is need for a judging group in the construction of attitude scales. Ferguson seems to believe that Murphy and Likert implied, as a result of obtaining a higher reliability coefficient with the 1 to 5 method of scoring than with the customary Thurstone method of scoring, that they had demonstrated that the method of summated ratings does away entirely with the need for a judging group. He argues against this and bases his criticism on the following grounds: "Since the statements (used by Murphy and Likert in the above study) had been sifted through the sorting procedure (Thurstone's), it would seem unjustifiable to conclude that Likert's method did away with the need for a judging group. To test this point adequately one should compare scales constructed (independently of the Thurstone method) by the Likert technique with those constructed by the equal appearing interval method" (7, p. 52).

We are in complete agreement with this argument; therefore we find ourselves at a loss to understand the following statement of experimental design appearing in the same article:

"A more adequate test can be provided by rescaling items using Thurstone's method in scales constructed by Likert's technique. If

of the test taken as a whole. We feel that, at least for purposes of comparison, it is not valid to raise the coefficients for the Thurstone scales which are based upon equivalent forms of 20 to 22 items each. To do so would indicate the reliability to be expected from a Thurstone scale of 40 to 44 items, while in practice the scales generally used contain only half these numbers.

Likert's technique does away with the need for a judging group, the two methods of treating the statements should give the same result" (7, p. 52).

But this particular experimental design will not give a test of the two methods of scale construction; it is an investigation of where Likert-selected items will fall along the continuum posited by Thurstone or, stated somewhat differently, what Thurstone scale values will be attached to the particular items included in a particular Likert-type scale.

What Ferguson found by following this line of investigation was that Likert-selected items, when scaled according to the method of equal appearing intervals, failed to spread evenly over the scale continuum of Thurstone; the statements failed to represent all degrees of attitude but fell largely at the favorable and unfavorable ends of the scale with the middle categories neglected. Only one of the Likert-type scales which Ferguson attempted to scale by the Thurstone technique, an economic conservatism scale, gave a fairly even spread of items, and the correlation between the Thurstone and Likert methods of scoring this scale was .70.¹² Because of these findings, the failure of the Likert-selected items to spread evenly over the Thurstone continuum and the "low" correlation between the Thurstone and Likert methods of scoring the one scale that did, Ferguson believes that he has successfully demonstrated "that Likert's technique for the construction of attitude scales does not obviate the need for a judging group" (7, p. 57).

We cannot agree with this conclusion. What has been demonstrated, as we pointed out earlier, is that Likert-selected items do not necessarily fall at equally-spaced intervals along the theoretical continuum posited by Thurstone and Chave. That they do not may be of theoretical interest, but has little bearing upon the practical problem of whether or not there is need for a judging group. This question can only be answered in terms of whether or not scales constructed *independently* by each of the two methods will yield comparable scores, i.e., if an individual is a standard distance above the mean on one scale, he will be a comparable distance above the mean on the second.

It might seem that the correlation of .70 between the Thurstone and Likert method of scoring the economic conservatism scale would bear upon the problem. But this correlation is biased in that Ferguson failed to give the Thurstone method a fair trial, i.e., he limited the Thurstone scale to the items already selected by Likert's technique. Nor can we accept the correlations of .75 and .81 (corrected for attenuation) which

¹² Assuming that the reliability coefficient of this scale when scored by the Thurstone method is approximately that obtained when the scale is scored by the Likert method (reported by Rundquist and Sletto as .85), and correcting for attenuation, the correlation would be .82.

Murphy and Likert report for their Internationalism Scale and scores on the Thurstone-Droba War Scale. These correlations also fail to do justice to the question of whether comparable results can be obtained with independently constructed Thurstone and Likert scales since it is possible that the attitudes under consideration are not the same.

Comparative Study of the Two Methods

A valid comparison of the Thurstone and Likert techniques, we believe, must start with an original set of items, not with items already sifted by the Thurstone procedure and then scored by Likert's method, and not with items sifted by the Likert procedure and then scaled by the Thurstone technique. We believe also that the same group of subjects should be used in the construction of the two scales, but that the steps for each method should be carried through independently. To carry out this comparison, we used the original statements of opinion used by Thurstone and Chave in the construction of their scale designed to measure attitude toward the church.

Subjects used in the construction of the scale were 72 members of an introductory psychology class at the University of Maryland.¹³ Half the class, selected at random, was asked to judge the degree of favorableness or unfavorableness expressed by the statements in accordance with the Thurstone method, while the other half of the class was requested to give Likert-type responses to the same statements. Two days later the procedure was reversed; the first half of the class gave Likert-type responses to the statements, while the other half gave Thurstone-type responses. The Seashore and Hevner (17) method of rating items was used instead of the Thurstone and Chave procedure of sorting items into piles.¹⁴

In constructing the Thurstone scale, tabulations were made indicating the number of judges who placed each item in each of the categories. From these data accumulative proportions were determined and ogives constructed for each item. Scale values of the items were found by dropping a perpendicular to the baseline of scale values at the point where the curve crossed the 50 per cent level.¹⁵ Q values were determined in a

¹³ Although Thurstone and Chave used a much larger group of subjects, subsequent research (14) indicates that groups as small as 25 or 50 can be used to obtain scale values of items and that these values are very similar to those obtained with larger groups.

¹⁴ Seashore and Hevner found that a technique of asking judges to rate statements on a scale instead of requesting that they sort the statements into piles yielded results which correlated very well with those obtained by Thurstone's original sorting procedure. See also the study by Ballin and Farnsworth (2).

¹⁵ The correlation coefficient between our scale values and those obtained by Thurstone and Chave 15 years earlier was .95. Our Q values, however, tended to differ considerably, the correlation coefficient being only .18.

similar fashion by dropping perpendiculars at the 25th and 75th per cent levels, the Q value being the scale distance between these two points.

Items were selected for two "equivalent" forms of the scale, each form containing 20 items. Selection was made on the basis of Thurstone's informal standards, Q values, and scale values. Insofar as possible, the final scales contained items with low Q values and with scale values which were spread along the entire scale at relatively equally-spaced distances. Since only a few items, however, were found to have scale values near the center of the continuum and, at the same time, low Q values, this was not entirely possible.

In constructing the Likert scale, a total score for each subject was found by summing the weights of responses given for each of the items. The upper and lower 10 subjects in terms of total scores served as the groups for applying the criterion of internal consistency. Since many of the original items would not meet Likert's *a priori* screening standards, we thought it possible that total scores determined in part by these items would include in the criterion groups individuals who might otherwise not be represented, i.e., if these items had not been used in the scoring. Thurstone and Chave, we might emphasize, had included various ambiguous items in the original set in order to test Q as a means of statistically determining ambiguity. Total scores, therefore, were first determined by excluding those items which we felt did not meet Likert's criteria. Total scores were then found with these items included. Since we found that the criterion groups would contain essentially the same subjects using either score, the total scores based on all of the items were used.

Twenty-five items, all with a mean difference between the two criterion groups of 1.8 or higher, were selected for the final Likert scale. Approximately half of these items were weighted 5 for a strongly agree response and half were weighted 1 for a strongly agree response. Of the 25 items selected for the Likert scale, 3 were also used in Form A of the Thurstone scale and 2 were used in common with Form B.

Reliability and Comparability

To obtain data on the reliabilities of the scales and to find out the relationship existing between scores on the *independently* constructed Likert and Thurstone scales, members of another introductory psychology class and an applied psychology class at the University of Maryland were tested. One group of subjects was presented with the Thurstone scales followed by the Likert scale; for the second group of subjects the order of presentation was reversed. There were 80 subjects altogether, each group containing approximately half of this number.

The reliability coefficient for the Likert scale of 25 items was .94. This coefficient compares favorably with those usually reported for scales

constructed by this method. The reliability coefficient for the equivalent forms of the Thurstone scales of 20 items each was .88. This is comparable to the reliability coefficients of .85 and .89 which Thurstone and Chave originally reported for scores on their Form A and B for two different groups of subjects (18, p. 66).

The correlation coefficient between scores on the Likert scale and Form A of the Thurstone was .72, which, when corrected for attenuation, becomes .79. On the other hand, the correlation between the Likert scale and Form B of the Thurstone was .92. When corrected for attenuation the coefficient indicates a perfect relationship. Unfortunately, we have no way of knowing which of these two coefficients is more representative of the "true" relationship existing between scores on independently constructed Likert and Thurstone scales in general. But the coefficient of .92 between the Likert and one of the Thurstone scales is surely sufficiently high to establish the fact that *it is possible* to construct scales by the two methods which will yield comparable scores. This is the question we set out to answer.

Summary and Conclusions

Now if we go back and examine the points on which we compared the Thurstone and Likert techniques of scale construction, we reach the following conclusions:

1. The evidence available indicates that the attitude of the judging group is not an important factor determining the scale values of items sorted by the Thurstone technique.¹⁶
2. Scales constructed by the Likert method will yield higher reliability coefficients with fewer items than scales constructed by the Thurstone method.
3. What evidence we do have seems to indicate that the Likert technique is less time-consuming and less laborious than the Thurstone

¹⁶ We are not satisfied with the evidence on this point. Would similar results obtain from judgments derived from those with sympathetic attitudes toward fascism and those violently opposed to fascism in the construction of a scale measuring attitude toward fascism? And in the case of communist sympathizers and non-communists in the construction of a scale measuring attitude toward communism? When social approval or disapproval attaches to a favorable or unfavorable attitude toward an issue, different scale values might result from groups with differing attitudes. An individual with a highly generalized unfavorable attitude toward fascism, for example, might scale an item such as: "Superior races are justified in dominating inferior races by force" as very favorable toward fascism. But would "native fascists" tend to scale it toward the same end of the continuum? The research so far, it seems to us, also neglects the related problem of *ego-involved* attitudes and the bearing they might have upon scale values of items.

technique. But additional research is needed on this point and should be based on carefully kept time records.

4. It is true that Likert-selected items tend to be those which would fall at one or the other extreme on the Thurstone continuum, if scaled according to the Thurstone technique. But the implication of this finding is more theoretical than practical as far as the need for a judging group is concerned. The important problem is whether scores obtained from the two differently constructed scales are comparable and the evidence at hand indicates that they are. As far as we can determine there is nothing of a practical nature to indicate that a judging group, in the Thurstone sense, is a prerequisite for the construction of an adequate attitude scale.

Received December 18, 1944.

References

1. Albig, W. *Public opinion*. New York: McGraw-Hill, 1939.
2. Ballin, M., and Farnsworth, P. R. A graphic rating method for determining the scale values of statements in measuring social attitudes. *J. soc. Psychol.*, 1941, **13**, 323-327.
3. Bird, C. *Social psychology*. New York: Appleton-Century, 1940.
4. Edwards, A. L. Unlabeled fascist attitudes. *J. abnorm. soc. Psychol.*, 1941, **36**, 575-582.
5. Ferguson, L. W. The influence of individual attitudes on construction of an attitude scale. *J. soc. Psychol.*, 1935, **6**, 115-117.
6. Ferguson, L. W. The requirements of an adequate attitude scale. *Psychol. Bull.*, 1939, **36**, 665-673.
7. Ferguson, L. W. A study of the Likert technique of attitude scale construction. *J. soc. Psychol.*, 1941, **13**, 51-57.
8. Hall, O. M. Attitudes and unemployment. *Arch. Psychol.*, N. Y., 1934, No. 165.
9. Hinckley, E. D. The influence of individual opinion on construction of an attitude scale. *J. soc. Psychol.*, 1932, **3**, 283-296.
10. LaPiere, R. T., and Farnsworth, P. R. *Social psychology*. New York: McGraw-Hill, 1942.
11. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.*, N. Y., 1932, No. 140.
12. Murphy, G., and Likert, R. *Public opinion and the individual*. New York: Harper, 1937.
13. Murphy, G., Murphy, L. B., and Newcomb, T. M. *Experimental social psychology*. New York: Harper, 1937.
14. Nystrom, G. H. The measurement of Filipino attitudes toward America by use of the Thurstone technique. *J. soc. Psychol.*, 1933, **4**, 249-252.
15. Pintner, R., and Forlano, G. The influence of attitude upon scaling of attitude items. *J. soc. Psychol.*, 1937, **8**, 39-45.
16. Rundquist, E. A., and Sletto, R. F. *Personality in the depression*. Minneapolis: Univ. Minnesota Press, 1936.
17. Seashore, R. H., and Hevner, K. A time-saving device for the construction of attitude scales. *J. soc. Psychol.*, 1933, **4**, 366-372.
18. Thurstone, L. L., and Chave, E. J. *The measurement of attitude*. Chicago: Univ. Chicago Press, 1929.

The Economy of Item Analysis with the IBM Graphic Item Counter *

Lt. Walter J. McNamara, U.S.N.R., and
Lt. Ellis Weitzman, U.S.N.R.

Central Examining Board, Naval Air Training Command, Pensacola, Florida

The Graphic Item Counter, an attachment for the IBM test scoring machine, has come into rather extensive use during recent years. This device is designed to provide item analysis data of objective test questions by mechanical means. By means of the attachment, it is possible to print in graphic form the responses to as many as 90 questions, for a maximum of 115 answer sheets. This attachment will also provide the necessary data for questionnaire analysis, or any type of response-counting, provided the original records are in the form of marks in particular positions on machine-scoreable answer sheets.

The item counter is equipped with a plugboard which has one position corresponding to each of the 750 possible response positions on the answer sheet. One end of each plugwire is connected to the response position to be counted, and the other end is plugged to one of the 90 counters in which it is desired to record the summation of marks in the particular answer space. Total counters are also provided which record the total number of sheets run through the machine. A switch is provided by means of which it is possible automatically to stop the machine at the end of a run of 100 papers. The plugboard may be wired to count the number of correct responses to 90 multiple-choice items at one time, or to record the number of responses to each of the choices of five-choice questions for 18 items at one run.

To analyze the marks on a group of answer sheets, each of the answer sheets is passed through the machine; the motor key is depressed once to put the answer sheet in position, and then the counter scans the sheet and makes the count of marks. As soon as this scanning is completed, the motor key is depressed again, and the sheet is automatically released and ejected from the machine. When the last answer sheet has been passed through the machine, a blank graphic item count record sheet and a piece of carbon paper are inserted in the machine, in much the same way as a

* The opinions or assertions contained herein are the private ones of the writers and are not to be construed as official or reflecting the views of the Navy Department or the naval service at large.

sheet of paper is inserted in a typewriter. By operating the print-start lever, the carriage automatically rolls the record sheet through the machine and prints on it a bar graph of the item count. The height of the bars for each item indicates the number of responses recorded for that item. It is then a simple matter to transfer the number of responses for each item, by means of a scale printed on each side of the record form, to individual item cards or other permanent record forms.

It is the purpose of this paper to present quantitative data relating to both the time and accuracy of the machine operations when compared with those done by hand. Since the clerical personnel is common to the tabulation of data derived from both machine and hand operations, this study is, basically, a comparison of the speed and accuracy of persons using these two methods of obtaining item analysis data.

The present inquiry was carried out at the Central Examining Board of the Naval Air Training Command. The Board prepares uniform tests which are administered at several naval training activities. At these training activities, students are given periodic objective tests in a number of subject matter fields—e.g., Principles of Flying, Aircraft Engines, Aerology, etc.—the results of which are analyzed by the Board and reported upon to the cognizant authorities.

This study is based upon analyses of answer sheets used exactly as they were received from the training activities, i.e., without any remarking of the sheets. In any situation involving many instructors administering tests to large groups, it is to be expected that not all of the students will mark all papers perfectly. Students sometimes record responses only to later erase them poorly before substituting other marks. Other examinees, despite instructions to the contrary, insist upon using hard lead pencils, or else mark too lightly. Greater preciseness of machine item analysis results if clerks scan answer sheets, remarking noticeably inadequate marks before running them through the machine. In this particular study, however, papers were not remarked, a fact which should be taken into consideration when viewing the resulting data.

To obtain a measure of the accuracy of the Graphic Item Counter itself, it would be necessary to use groups of perfectly marked papers. The present study, however, does not deal with the mechanical accuracy of the machine itself, but with the analysis of results obtained from run-of-the-mill answer sheets.

In the present study, groups of 100 test papers (answer sheets) were used for each test included, since this gives percentage figures without conversion. All tests involved in the study were made up of five-choice questions. Since an accurate count of the number responding to each choice (students being requested to respond to all questions) should total

100 cases for each question, the extent to which the tabulations for each question equal 100 is an index of accuracy. The first check consisted of seeing to what extent the tabulations for each item approximate 100. Table 1 shows the results for tabulations from the machine record for 320 test items in four different Physics tests, 710 test items from six different tests in Principles of Flying, and 640 test items from seven tests in Operation of Aircraft Engines. The table gives the percentage of items totalling 100, as well as the percentages above and below it. For each of the test items there were 100 answer sheets (or examinees).

As seen in Table 1, 34.4 per cent of the items showed a total, for the choice counts, of 100. Within the range of plus-or-minus one per cent

Table 1
Totals of the Per Cent Figures for the Five Choices

Percentage Totals	Four Physics Tests		Six Prin. of Flying Tests		Seven Engines Tests		All Tests Combined	
	No. Items	%	No. Items	%	No. Items	%	No. Items	%
96 and below	12	3.7	43	6.0	18	2.9	73	4.4
97	13	4.1	56	7.9	48	7.5	117	7.0
98	54	16.9	120	16.9	96	15.0	270	16.2
99	92	28.8	208	29.3	147	22.9	447	26.8
100	104	32.5	209	29.4	262	40.9	575	34.4
101	37	11.6	42	5.9	48	7.5	127	7.6
102	4	1.2	9	1.3	15	2.3	28	1.7
103	2	0.6	2	0.3	5	0.8	9	0.5
104 and above	2	0.6	21	3.0	1	0.2	24	1.4
Totals	320	100.0	710	100.0	640	100.0	1670	100.0

were found 68.8 per cent of the items. A range of plus-or-minus two per cent includes 86.7 per cent of the cases, while within a range of plus-or-minus three per cent are found all but 5.8 per cent of the cases.¹ It is also evident that in tabulation from machine records there is a much greater tendency to go below rather than above the 100 point. The percentage of totals below 100 is 54.4, while the percentage of totals above 100 is only 11.2. This probably results from the fact that there are very few cases in which there are extra marks on the answer sheets which result in a total count of more than 100. On the other hand, it is more often the case that students fail to answer a particular question (despite directions to answer all questions), or else mark so poorly that the machine does not register them.

In order to see the extent of agreement between two tabulations made

¹ In actual practice, the board rechecks all ~~cases~~ that vary to any appreciable degree.

from machine data, sets of examination papers were run through, by regular clerks, on two different machines (without remarking the papers). These two runs were made to see what agreement exists, *for the correct choices only*, between the two separate machine runs (using two machines and two clerks). Table 2 shows the direction and degree of divergence of the second run when compared with the first.

From Table 2, it is evident that in tabulations from machine analysis a tolerance of plus-or-minus one per cent will give tabulations which are accurate in 71.7 per cent of the cases. (Agreement of two separate tabulations of machine data is in this case the assumed criterion of accuracy.)

Table 2

Percentage Differences Obtained from a Second Run (Correct Choice Only)

Percentage Differences	Four Physics Tests		Six Prin. of Flying Tests		Seven Engines Tests		All Tests Combined	
	No. Items	%	No. Items	%	No. Items	%	No. Items	%
+4 and above	8	2.5	4	0.6	12	1.9	24	1.4
+3	9	2.8	48	6.7	4	0.6	61	3.7
+2	32	10.0	71	10.0	22	3.5	125	7.5
+1	40	12.5	99	13.9	144	22.5	283	17.0
0	96	30.0	244	34.4	258	40.3	598	35.8
-1	60	18.7	156	22.0	100	15.6	316	18.9
-2	33	10.3	69	9.7	59	9.2	161	9.6
-3	12	3.8	12	1.7	18	2.8	42	2.5
-4 and below	30	9.4	7	1.0	23	3.6	60	3.6
Totals	320	100.0	710	100.0	640	100.0	1670	100.0

A tolerance of plus-or-minus two per cent includes 88.8 per cent of the cases studied, while a tolerance of plus-or-minus three per cent includes 95.0 per cent of the cases. Five per cent of the cases show deviations as great as, or greater than, plus-or-minus four per cent.

Whether or not the obtained deviations from "accuracy" are too great for a particular testing program depends upon at least two factors. The first of these is the exactness of measurement involved in the testing itself. This, in turn, is a function of the validity and reliability of the tests administered and the adequacy of the sample used in making analyses. The second factor to be considered is the exactness required in the analysis data. This factor is generally the determining one in most situations. (Certainly, the reporting of the results of testing should not imply any greater precision than is noted in the analysis data themselves.) Although the matter of the required degree of accuracy of item analysis must be dealt with in terms of the specific needs of a specific testing pro-

gram, the writers feel that the degree of accuracy indicated above is probably adequate for most programs.

Accuracy of Tabulations from Machine Compared with Tabulations by Hand

For purposes of comparing the accuracy of item analysis tabulations made by hand with those made when using the Graphic Item Counter, six 40-item tests were selected. These six tests were composed of two each in the subjects Aerology, Engines, and Principles of Flying. A check was made for each of the five choices in the six 40-item tests, a total of 1,200 choices. As above, the data used for each item were for 100 answer sheets (or examinees). Accuracy was determined by first making a tabulation from the Graphic Item Counter records, then making a hand tabulation. Whenever the two tabulations were in agreement,

Table 3
Number of Cases of Incorrect Tabulations

Type of Tabulation	Errors		Total Errors	
	± 1	± 2	No.	%
Hand	161	92	253	71.9
Machine	50	26	76	21.6
Both	15	8	23	6.5
Totals	226	126	352	100.0

it was judged that the tabulation was correct. Whenever discrepancy existed between the two separate tabulations, an additional tabulation was made by hand. Agreement with the second hand tabulation was the basis for determining whether the machine tabulation or the hand tabulation had been inaccurate.

The hand tabulations were in perfect accord for all five choices with those which had been made with the Graphic Item Counter in the case of 40 of the 240 test questions. Of the remaining 200 items, 38 items showed agreement on the correct choices. This gives 78 cases (32.5 per cent) in which the two tabulations agreed with respect to the percentages, selecting the correct choices. (This may be compared with the 34.4 per cent of perfect agreement in Table 1.) Of the total of 1,200 choices (240 questions), there was agreement between hand and machine tabulations on 848 choices. The two tabulations were at variance in 352 of the 1,200 choices. Table 3 shows the results obtained when the second hand tabulation was made to determine which of the two previous (one hand and one machine) tabulations had been in error, as well as the size of the error.

As seen in Table 3, of the 352 cases of hand and machine disagreement, the machine check was off in 76 cases, the hand tabulation was off in 253 cases, while in the remaining 23 cases both hand and machine tabulations were inaccurate. Of the total of 352 cases of error in tabulating responses, 118 deviated two or more per cent. Of these 118 cases deviating by two or more per cent, hand tabulations were responsible for 92 of the 118 cases. In percentage terms, the hand tabulations were responsible for 71.9 per cent of the errors, the machine tabulations for 21.6 per cent, while 6.5 per cent of the errors occurred in both types of tabulations.

In terms of the total number of responses in the six tests, 1,200, the tabulations made from machine data were in error in 6.3 per cent of the 1,200 choices; the hand tabulations were in error in 21.1 per cent of the 1,200 cases; and both hand and machine tabulations were in error in 1.9 per cent of the 1,200 cases. From this, it appears that when tabulations are made for Graphic Item Counter data, item analysis tabulations are in error only about one-third as often as when tabulations are made by hand exclusively.

In terms of errors in excess of plus-or-minus one per cent, machine tabulations showed such errors in 2.2 per cent of the 1,200 cases, while the hand tabulations showed such errors in 7.7 percent of the 1,200 cases. Again, the machine tabulation is inaccurate to a degree of plus-or-minus two per cent only one-third as often as is the hand tabulation. Put in positive terms, the tabulations from the machine record are within two per cent of accuracy in 97 per cent of cases, while the hand tabulations are within two per cent of accuracy in 92 percent of the cases.

Time Required for Hand Tabulations Compared with that for Machine Tabulations

The investigation of the accuracy of machine versus hand item analysis presented above also resulted in some interesting comparative data on the time required for the two types of tabulations. Table 4 shows the time requirements for the hand and the machine operations referred to in the preceding pages.

The time data for the hand tabulations show considerable variation. This is due, no doubt, to the fact that four clerks prepared the first four sets of hand tabulations, a fifth preparing the last two listed. The tabulations made from the machine item analysis were made by the clerk who had prepared the hand tabulations for Flying tests A and B (the two requiring the least amount of time).

For making the machine item analysis, 30 minutes were required for the clerk to wire the plugboards for counting the 200 choices in the 40-item, five-choice tests. Two plugboards were used. The data for the first 18 items were obtained by using the first plugboard, the data for

Table 4
Tabulation Time Record

Test	Hand-Tabulation Time	GIC-Tabulation Time
Aerology A	6 hr. 40 min.	32 min.
Aerology B	6 hr. 30 min.	30 min.
Engines A	5 hr. 25 min.	30 min.
Engines B	8 hr. 0 min.	30 min.
Flying A	5 hr. 20 min.	29 min.
Flying B	4 hr. 45 min.	31 min.
Wiring time		30 min.
Totals	36 hr. 40 min.	3 hr. 32 min.

items 19-36 by using the second plugboard. To obtain a count on items 37-40, the wires on the second plugboard, for the last four items (33-36), were merely moved down four rows. It should be noted that because the tests used had 40 five-choice items it was necessary to make a third run on the machine to get the data on the last four items. If four-choice items for a 40-item test had been used, the necessary data could have been obtained in two runs. Similarly, for five-choice items it is possible with three runs through the machine to obtain data on as many as 54 items. Thus, in making comparisons with hand tabulations, the differences are not in this instance the maximum which are possible.

The average time required to run the six tabulations using Graphic Item Counter item analysis was about 35 minutes, including wiring time. The average time for the hand tabulations was about six hours. The average time required for the hand tabulation for the clerk who ran the machine analysis was about five hours as compared with an average time of about 35 minutes per machine tabulation. In round figures, the clerk making the speediest hand tabulations required over eight times as long for hand analyses as was needed for tabulations made with the Graphic Item Counter.

Summary

1. Item analyses made by clerks using the Graphic Item Counter attachment of the IBM test scoring machine are more accurate than those made by hand tabulation alone.

2. Item analysis tabulations made by clerks using the Graphic Item Counter are sufficiently accurate to meet the requirements of most testing programs.

3. Item analysis tabulations made by hand take more than eight times as long as do those made with the Graphic Item Counter.

Received March 24, 1945.

The Interrelationship of Visual Acuity at Different Distances*

William James Giese

Division of Education and Applied Psychology, Purdue University

In present day employment methods it is almost universally standard practice to administer a test of visual acuity. In many companies this test consists of the standard Snellen chart administered at a distance of twenty feet. The tacit assumption is made that normal vision, 20/20 rating on the Snellen test, is "good" vision; the practice is that for jobs requiring "good" vision personnel should rate 20/20 on the test. In practice this method of personnel allocation is often used on the basis of arbitrary judgment, and when validation studies are made it is sometimes found that this method of vision testing is far from satisfactory. Indeed, in some instances the use of a single distance acuity test actually eliminates the potentially successful workers and selects those visually unfitted for the job. Tiffin and Wirt in their investigation of visual acuity and hourly production of hosiery loopers found a correlation of approximately — .60 between visual acuity measured at twenty feet and production.¹ The work of these loopers was at a visual distance of only eight inches which presents quite a different visual task than passing a test of visual acuity at twenty feet. From the point of view of the placement of industrial employees in jobs for which they are visually qualified it is important to know what the relationships are between visual skill tested at various distances. In addition, it is desirable to know the variation of average acuity for different distances as well as the variation of the spread of individual differences about the mean acuity for different distances.

The present research was undertaken to provide an answer based on experimental evidence to the following three main points:

- (1) To determine experimentally the relation between visual acuity at various distances.
- (2) To determine the average visual acuity in terms of minute angles at various distances.
- (3) To determine the extent of individual differences in visual acuity at various distances.

* This article is based on the author's thesis of the same title submitted to the faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, October, 1945. The thesis was directed by Dr. Joseph Tiffin.

¹ Joseph Tiffin and S. E. Wirt. Near vs. distance acuity in relation to success on close industrial jobs. *Trans. Amer. Acad. Ophth. and Otolaryng.* (June 1944), Suppl. pp. 6-16.

Procedures

A multiple choice checker test for visual acuity² was chosen as the instrument for measuring visual acuity (see Figure 1). This visual target is so proportioned that for "normal" vision at any distance the individual checkers in the test square subtend a visual angle of one minute while the individual checkers in the four remaining control squares subtend a visual angle of 12 seconds. The checker design in the control squares is so fine that for individuals of exceedingly high

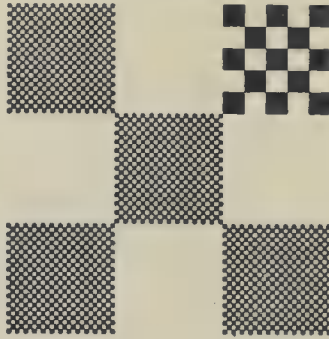


FIG. 1. Checker board design.

acuity they blend into a neutral grey. All of the visual targets were mounted in the center of an eight by eight inch grey card which matched the grey of the control squares. These cards were presented to the subjects in a specially designed box which kept such factors as illumination and ocular distance constant. When this design presents a visual angle which is below a subject's visual acuity to resolve, the test square blends into a neutral grey matching the control squares leaving the subject's response one controlled by chance.

The checker design was used instead of the more customary letter chart because of the following limitations of the letter chart:

- (1) Measures in addition to visual acuity readability of letters.³
- (2) Does not lend itself readily to safe reduction especially for short distances.
- (3) Adequately differentiates only substandard levels of acuity while at standard and superior levels it differentiates only grossly.⁴
- (4) The reliability and validity of the chart as a whole are difficult to establish since these factors are different for each letter in the chart.⁴

² Developed by the Scientific Bureau of the Bausch & Lomb Optical Co., Rochester, New York.

³ J. P. S. Walker. Test type. *Brit. J. Ophthal.* (1942), 26, 556-559.

⁴ Joseph Tiffin, *Industrial psychology*. New York: Prentice-Hall, Inc., 1944, pp. 128-129.

Each of these four limitations of the letter chart are strong reasons for the use of the multiple choice checker visual acuity test design since it has been carefully designed so that the disadvantages of the letter test have been eliminated or greatly reduced. The apparatus was set up so that eight levels of visual acuity could be obtained at the distances of .20, .25, .33, .40, .50, 1.00, 5.00, and 10.00 meters which require a diopter change in focal power from infinity of 5, 4, 3, 2, 1, .2, and .1 respectively. The room in which the apparatus was installed was large and well ventilated and fitted with black out shutters which were adjusted so the illumination in the vicinity near the apparatus was under one foot candle. The illumination on the visual target was eight foot candles which is one foot candle more than is recommended by Ferree and Rand as the optimum illumination for the purpose of obtaining measurement of visual acuity.⁵ The purpose and method of taking the test were explained through standard instructions to each subject. Half of the subjects started at the near distance and worked back to the far distance while half started at the far distance and worked down to the near distance. At each distance the subject started with the target with the smallest visual acuity rating (the largest visual angle) which was increased by .2 visual acuity rating steps until he made a wrong judgment. The subject was then presented with the previous target on which he was asked to make four judgments. If he failed to make four correct judgments he was then presented with the target of the next larger visual angle and again asked to make four judgments. This procedure was continued until the subject was able to make four successive correct judgments. The subject's visual acuity rating for any distance was, then, the smallest target on which he could make one original and four later successive correct judgments. Since there are four alternatives for each of the five judgments, a subject has only one chance in 1024 to make five correct judgments when he cannot see the location of the checkers. Four hundred subjects were tested for their visual acuity by this method, and eighty-nine additional subjects were tested with the only change in the procedure a retest at each distance for the purpose of determining the reliability of the tests.⁶ All of the subjects had previously been screened by a 20/20 standard at twenty feet on a letter chart. The median age of the group was 19 with none under 17, but some as old as 36.

⁵ C. E. Ferree and G. Rand. New ideas in eye testing. *Personnel J.* 1939, 18, 13-20.

⁶ A complete detailed description of the experimental procedure is given in W. J. Giese, "The interrelationship of visual acuity at different distances," a thesis submitted to the faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology, October, 1945.

Results

Reliabilities. Table 1 summarizes the statistics concerned with the test-retest reliabilities of the multiple choice checker acuity test for various distances. With these subjects acuity measured at 5.00 and 10.00 meters is more reliable than acuity measured at shorter distances. Probably the reason for this is that fluctuations in accommodation and convergence are more likely to occur in near than in far vision tasks. Such fluctuations could account for the somewhat lower reliabilities of the near vision tests. Obtained Pearson Product Moment coefficients of correlation corrected for grossness of grouping appear under column headed $r_{x_1'x_2'}$ which show that since the lower reliabilities of the near tests are not due to variation in grouping there are factors other than

Table 1
Test-Retest Reliabilities of the Multiple Choice Visual Acuity
Checker Test at Various Distances

Distance in Meters	$r_{x_1x_2}$	$r_{x_1'x_2'}$	$e_{x_2'x_1'}$	$\eta_{x_2x_1}$	$e_{x_2x_1}$	X^2	P	$e_{x_1'x_2'}$	$\eta_{x_1x_2}$	$e_{x_1x_2}$	X^2	P
.20	.87	.92	.92	.88	.87	5.90	.32	.91	.87	.86	2.81	.73
.25	.72	.77	.81	.78	.76	18.49	.0024	.80	.75	.74	8.49	.19
.33	.82	.87	.90	.86	.85	19.56	.0017	.92	.83	.82	3.65	.44
.40	.80	.84	.89	.86	.85	30.36	.00001	.85	.83	.81	10.56	.06
.50	.80	.84	.85	.82	.81	6.27	.29	.84	.82	.80	5.55	.36
1.00	.82	.87	.87	.83	.82	3.21	.67	.86	.83	.81	1.72	.87
5.00	.91	.97	.98	.92	.92	8.74	.12	.97	.92	.91	4.09	.54
10.00	.93	1.00	1.00	.93	.93	3.07	.55	1.00	.93	.93	3.55	.47

coarse test step intervals which reduce the reliability. If it were possible to decrease the size of the test step intervals without altering the fatigue factor one should not expect the reliability coefficients to be greater than those listed under column $r_{x_1'x_2'}$. Through a comparison of column $r_{x_1x_2}$ with column $r_{x_1'x_2'}$ the conclusion can be drawn that although the increase in reliability would be desirable it would not have been feasible to attempt to achieve it through finer test step intervals especially since they would have greatly increased the length of time for test administration. For all but three of the near distances a linear regression line gives a satisfactory fit, and it is only when *eta* is computed in the initial test scores is there statistical significance to the fact that some line other than a straight one would be a more satisfactory fit.⁷ An inspection of

⁷ Chi Square = $(N - K) \frac{\eta_{yx}^2 - r_{xy}^2}{1 - \eta_{yx}^2}$. C. C. Peters and W. R. Van Voorhis. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill Book Co., Inc., 1941.

the scattergrams showed that this was caused by a selective regression to the mean on the retest. The cases with very high acuity on the initial test, as a group, did more poorly on the retest, but the cases with low acuity ratings, as a group, did not shift one way or the other on retest. However, by a comparison of columns $r_{x_1x_2}$, $\epsilon_{x_1x_2}$, $\epsilon_{x_2x_1}$, it can be seen that although the non-linearity of the reliabilities of the distances of .25, .33, and .40 meters has high statistical significance, the difference in degree of relationship expressed by a linear or curvilinear coefficient is so small that they have no practical import. Another factor that should

Table 2

Comparison of Means and Standard Deviations on Initial and Retest

Distance	Initial		Retest		Diff. I - R	SE Diff.	CR	P
	Mean	SE	Mean	SE				
.20	.89±.0225		.89±.0218		.00		.00	1.00
.25	.98±.0213		.98±.0210		.00		.00	1.00
.33	1.19±.0232		1.18±.0210		.01	.018	.56	.42
.40	1.27±.0256		1.27±.0256		.00		.00	1.00
.50	1.29±.0254		1.26±.0249		.03	.011	2.73	.006
1.00	1.49±.0365		1.51±.0359		-.02	.022	.91	.37
5.00	1.60±.0315		1.61±.0316		-.01	.013	.77	.45
10.00	1.54±.0292		1.56±.0293		-.02	.011	1.82	.07

Distance	Initial		Retest		Diff. I - R	SE Diff.	CR	P
	Sigma	SE	Sigma	SE				
.20	.2121±.0159		.2053±.0154		.0068	.0080	.85	.40
.25	.2012±.0151		.1981±.0148		.0031	.0112	.30	.76
.33	.2190±.0164		.1981±.0148		.0209	.0094	2.22	.03
.40	.2416±.0181		.2419±.0181		-.0003	.0115	.03	.98
.50	.2396±.0180		.2349±.0176		.0047	.0118	.40	.70
1.00	.3444±.0258		.3391±.0254		.0053	.0154	.40	.70
5.00	.2968±.0222		.2982±.0224		-.0020	.0095	.21	.83
10.00	.2751±.0206		.2767±.0207		-.0016	.0077	.21	.83

be isolated is the effect of learning. Table 2 lists the means on initial test and retest as well as the difference between means and the probabilities that the difference could have arisen by chance. Similar data is presented for the standard deviations. There is only one P value, for the .50 meter distance, which indicates a statistically significant shift in the means, and it is in the direction of negative learning. It is, however, a rather small absolute shift, .03, when compared to the step intervals of the tests, .20. With the exception of this small negative shift there is no learning as revealed by shifts in the means. A similar analysis was made for shifts in the standard deviations in which the

smallest P value was .03 and all but two of the P values were .70 or above. Also, four of the shifts were in the direction of smaller standard deviations on the retest while four shifted in the direction of larger standard deviations on the retest. The P value of .03 was for a smaller standard deviation on retest for the .33 meter distance which was one of the distances for which a straight line had a very low probability for the best fitting regression line.

In summary, the statistics on the test-retest reliabilities show that the tests have adequate reliability, and that the means and standard deviations are stable from test to retest.

The Relationship of Visual Acuity at Different Distances. The greater the difference in diopters of accommodation required in the vision tests the less the degree of relationship between them. Stated in terms of

Table 3
Intercorrelation of Visual Acuity at Various Distances
(Obtained Pearson Product Moment Coefficients of Correlations)

Distance	.20	.25	.33	.40	.50	1.00	5.00	10.00
.20	.87*							
.25	.50	.72*						
.33	.46	.43	.82*					
.40	.41	.39	.52	.80*				
.50	.39	.42	.42	.49	.80*			
1.00	.29	.27	.35	.50	.50	.82*		
5.00	.17	.27	.34	.38	.45	.53	.91*	
10.00	.33	.36	.37	.41	.41	.32	.56	.93*

* Reliability coefficients from a sample of 89 cases apart from the sample of 400 cases used in the intercorrelations.

test distances this means that the greater the difference in the test distance between two vision tests the smaller the relationship between the acuity measures. Table 3 gives the obtained correlation matrix from which it is clear that all of the interdistance correlations are substantially smaller than the coefficients of reliability. Correlations between adjacent distances have a median of .52 while the median reliability correlation is .82. Table 4 summarizes the effect of differences in distances and degree of relationship between tests.

Although the differences which appear in the reliabilities for various distances could not radically change the pattern of the results, they might partially obscure or distort them. The range in the reliabilities is from .72 to .93 so some of the differences in interdistance correlations could be due to the fact that the reliabilities varied. Table 5 presents the

Table 4

Median Correlations by Amount of Separation Between Distances
(Obtained Pearson Product Moment Coefficients of Correlation)

Reliability	Adjacent					
0 Distances away	1 Distance away	2 Distances away	3 Distances away	4 Distances away	5 Distances away	6 Distances away
.82	.52	.44	.41	.36	.29	.26

matrix of correlation in Table 3 corrected for attenuation with the results, of course, that all of the correlations are somewhat higher. However, the essential relationship between the correlations not only has the same structure but this structure is more clear. For instance, the range of correlations between adjacent distances is only from .60 to .64, but in

Table 5

Intercorrelation of Visual Acuity at Various Distances
(Obtained Pearson Product Moment Coefficients of Correlation
Corrected for Attenuation)

Distance	.20	.25	.33	.40	.50	1.00	5.00	10.00
.20								
.25	.63							
.33	.55	.60						
.40	.49	.51	.64					
.50	.47	.55	.52	.61				
1.00	.33	.34	.42	.60	.60			
5.00	.19	.33	.39	.45	.53	.60		
10.00	.37	.44	.42	.48	.48	.36	.61	

the uncorrected matrix the range was from .43 to .56. Table 6 presents the corrected median correlations for easy comparison to Table 4.

Table 6 shows the relationship between distances if the measures had perfect reliability. Even with perfect reliability the median correlation with an adjacent distance would be only .60. If the distances

Table 6

Median Correlations by Amount of Separation Between Distances
(Pearson Product Moment Coefficients of Correlation Corrected for Attenuation)

Reliability	Adjacent					
0 Distances away	1 Distance away	2 Distances away	3 Distances away	4 Distances away	5 Distances away	6 Distances away
1.00	.60	.52	.48	.43	.33	.31

of .40 and 5.00 meters are eliminated from the matrix in Table 5, the difference between distances would then represent one diopter change in focal power required with the exception of 1.00 to 10.00 meters which is .9 instead of 1.0. Although this reduces the number of correlations from which to draw a median, it has the advantage that the medians can be listed for different amounts of diopter change required, a more meaningful concept in terms of the visual task confronting the eye.

Table 7

Median Correlations by Amount of Diopter Change in Focal Power Required
(Pearson Product Moment Coefficients of Correlation Corrected for Attenuation)

Reliability no change	1 Diopter	2 Diopters	3 Diopters	4 Diopters	5 Diopters
1.00	.60	.51	.42	.38	.37

Table 7 clearly shows that the greater the diopter change in focal power required of the eye from one visual acuity test to another the less the degree of relationship between the two measures. Even one diopter of change required greatly reduces the relationship. For three diopters or more change the relationship between tests is very low. Once the visual task has been changed three diopters, it might as well be changed four or five since the relationship drops only minutely for the greater changes.

All of the interdistance correlations presented so far have been linear coefficients. Inasmuch as coefficients of curvilinear correlation might give a different structure to the results, a matrix of correlation ratios is presented in Table 8. Here again, the general pattern of the results

Table 8

Intercorrelation of Visual Acuity at Various Distances
(Correlation Ratios—Eta)

Distance	.20	.25	.33	.40	.50	1.00	5.00	10.00
.20	.88*							
.25	.51	.78*						
.33	.49	.44	.88*					
.40	.46	.42	.53	.86*				
.50	.44	.44	.42	.50	.82*			
1.00	.44	.31	.40	.55	.53	.83*		
5.00	.26	.31	.37	.44	.47	.55	.92*	
10.00	.34	.36	.40	.42	.42	.33	.57	.93*

* Reliability coefficients from a sample of 89 cases apart from the sample of 400 cases used in the intercorrelations.

Table 9

Median Correlation Ratios by Amount of Separation Between Distances

Reliability	Adjacent	2	3	4	5	6
0 Distances away	1 Distance away	Distances away	Distances away	Distances away	Distances away	Distances away
.87	.53	.44	.44	.39	.40	.31

remains the same, and a summary of the median relationship between differences in distances and amount of relationship, Table 9, shows no change in the effect noted with the linear coefficients with the exception that when the distances between the tests are great, the relationship is not reduced quite so much.

Table 10 shows that when *etas* are used as a measure of interrelationship a change of just one diopter in focal power required makes the relationships between tests very low. Even though there is a consistent downward trend of the relationships as the change in diopters of accommodation increases, additional changes beyond one in diopter requirements between tests do not attenuate intertest relationships more than minutely.

Table 10

Median Correlation Ratios by Amount of Diopter Change in Focal Power Required

Reliability	1	2	3	4	5
no change	Diopter	Diopters	Diopters	Diopters	Diopters
.85	.44	.43	.40	.35	.34

One objection to *eta* as a measure of relationship is that it is generally too large since chance variations may often be of such a nature as to reduce the variance with the columns as compared to the variance of the total distribution. This objection is particularly cogent when applied to data composed of few cases and separated into many columns. The interrelationship figures in this study are based on over 400 cases which should be adequate to minimize the effect of chance variation, especially since the number of columns is always 7 or less. In the test-retest data, however, the number of cases is only 89, though again, the number of columns is always 7 or less. Because of the objections to *eta* as a measure of correlation the correlations are also expressed by *epsilon*, a correlation ratio without bias. These appear in Table 11. The maximum decrease from the *etas* is .03 for the interrelationship ratios and .03 for the intra-relationship ratios. Again, the general structure of the results is similar. In the interest of completeness these correlation

Table 11
Intercorrelation of Visual Acuity at Various Distances
(Correlation Ratios without Bias—Epsilon)†

Distance	.20	.25	.33	.40	.50	1.00	5.00	10.00
.20	.87*							
.25	.50	.76*						
.33	.48	.44	.85*					
.40	.45	.41	.52	.85*				
.50	.41	.43	.42	.49	.81*			
1.00	.43	.30	.39	.55	.52	.82*		
5.00	.23	.29	.36	.43	.46	.55	.92*	
10.00	.32	.35	.39	.41	.42	.31	.56	.93*

† All of the above ratios reach the 1% level of significance.

* Reliability coefficients from a sample of 89 cases apart from the sample of 400 cases used in the intercorrelations.

ratios without bias are corrected for grossness of grouping and the resulting matrix is shown in Table 12. An inspection of this table shows that with only a few minor reversions, the relationship between acuity measures decreases as the difference in diopters of focal power increases.

Table 12
Intercorrelation of Visual Acuity at Various Distances
(Correlation Ratios without Bias—Epsilon corrected for Broad Categories)

Distance	.20	.25	.33	.40	.50	1.00	5.00	10.00
.20	.92*							
.25	.56	.81*						
.33	.52	.49	.90*					
.40	.50	.46	.58	.89*				
.50	.45	.49	.43	.54	.85*			
1.00	.47	.33	.42	.60	.54	.87*		
5.00	.25	.32	.39	.47	.50	.59	.98*	
10.00	.34	.38	.41	.44	.44	.33	.59	1.01*

* Reliability coefficients from a sample of 89 cases apart from the sample of 400 cases used in the intercorrelations.

The Relationship Between Mean Acuity and Distance. The mean visual acuity steadily becomes smaller as the test distance becomes smaller. Table 13 gives both the mean acuity and the standard deviation about the mean, and Figure 2 presents the same data graphically. The data show that for distances requiring two or more diopter change in focal power the spread of individual differences is significantly lower than those distances requiring one diopter or less. Also, from the dis-

tance of 1.00 meter down to .20 meter (1 diopter to 5 diopters) the mean steadily decreases as the distance decreases. This result agrees with the findings of Luckiesh and Moss in which they determined the acuity threshold of 10 subjects for the distances of .60, 1.20, and 2.80 meters and found that for each subject acuity, measured by the reciprocal of the visual angle in minutes subtended by the critical detail, increased with increase in distance between observer and stimulus.⁸ Comparing Table

Table 13

Visual Acuity: Means and Standard Deviations in Visual Decimal Obtained on the Multiple Choice Checker Test
(Obtained from 400 cases used for the intercorrelations)

Distance in Meters	Mean \pm SE	Standard Deviation \pm SE
.20	.95 \pm .0100	.199 \pm .0070
.25	1.05 \pm .0098	.194 \pm .0069
.33	1.28 \pm .0095	.190 \pm .0068
.40	1.37 \pm .0083	.166 \pm .0059
.50	1.39 \pm .0099	.198 \pm .0070
1.00	1.63 \pm .0142	.283 \pm .0101
5.00	1.61 \pm .0154	.307 \pm .0109
10.00	1.35 \pm .0153	.305 \pm .0108

13 with the tables of the correlations, no consistent relationship between any of the Pearson Product Moment coefficients of correlation (obtained, corrected for attenuation, or corrected for broad categories) or the three correlation ratios and the standard deviations is found. In addition, there is no relationship between variation in the means and variation in the size of the measures of correlation. One fact does stand out, however; the nearest and the two farthest distances have the highest reliabilities. Of the three highest sigmas for the test-retest data, two

Table 14

Differences in Means of the Near Point Tests from All Farther Point Tests Along with Their Standard Errors of the Difference

Distance	.20	.25	.33	.40	.50	1.00	5.00
.20							
.25	.10 \pm .0141						
.33	.33 \pm .0140	.23 \pm .0138					
.40	.42 \pm .0131	.32 \pm .0122	.09 \pm .0127				
.50	.44 \pm .0142	.34 \pm .0141	.11 \pm .0139	.02 \pm .0130			
1.00	.68 \pm .0177	.58 \pm .0174	.35 \pm .0173	.26 \pm .0166	.24 \pm .0175		
5.00	.66 \pm .0186	.56 \pm .0184	.33 \pm .0183	.24 \pm .0176	.22 \pm .0185	-.02 \pm .0211	
10.00	.40 \pm .0185	.30 \pm .0184	.07 \pm .0182	-.02 \pm .0176	-.04 \pm .0184	-.28 \pm .0211	-.26 \pm .0219

⁸ M. Luckiesh and F. K. Moss. The dependency of visual acuity upon stimulus distance. *J. opt. Soc. Amer.*, XXIII, pp. 25-29.

are for the distances of 5.00 and 10.00 which also have the two highest reliabilities. Although the sigma for the .20 meter distance is small, the distribution of acuity scores at this distance is substantially normal which should, partially at least, account for a somewhat higher reliability at this distance. The differences between the means are tabulated in Table 14 which shows that except for the 5.00 and 10.00 meters distances there is a consistent increase in the mean performance as the number of diopters of accommodation required decreases. But how many of these shifts have statistical significance? The critical ratios for the shifts are presented in Table 15, an inspection of which shows that all

Table 15
Significances of the Differences (Critical Ratios) in Table 14

Distance	.20	.25	.33	.40	.50	1.00	5.00
.20							
.25	7.09						
.33	23.57	16.67					
.40	32.06	26.23	7.07				
.50	30.99	24.11	7.91	1.54			
1.00	38.42	33.33	20.23	15.66	13.71		
5.00	35.48	30.43	18.03	13.64	11.89	.95	
10.00	21.62	16.30	3.85	1.14	2.17	13.27	11.87

of the shifts are significant except .50-.40, 10.00-.40, 10.00-.50, and 5.00-1.00 when the rule of thumb of a critical ratio of 3.00 or better for significance is used. Not only do the majority of the shifts have a very high statistical significance but they also have practical import when the shift of the total distribution is considered. In addition to a shift in the means, there is a difference in the spread of the distributions. Table 16 tabulates the differences in spread which shows, in general,

Table 16
Difference in Standard Deviations of the Near Point Test from All Farther Point Tests
Along with Their Standard Errors of the Difference

Distance	.20	.25	.33	.40	.50	1.00	5.00
.20							
.25	-.005 ± .0099						
.33	-.009 ± .0098	-.004 ± .0098					
.40	-.033 ± .0092	-.028 ± .0092	-.024 ± .0091				
.50	-.001 ± .0100	.004 ± .0099	.008 ± .0098	.032 ± .0092			
1.00	.084 ± .0124	.089 ± .0123	.093 ± .0123	.117 ± .0118	.085 ± .0124		
5.00	.108 ± .0131	.113 ± .0130	.117 ± .0130	.141 ± .0125	.109 ± .0131	.024 ± .0150	
10.00	.106 ± .0130	.111 ± .0130	.115 ± .0129	.139 ± .0124	.107 ± .0130	.022 ± .0149	-.002 ± .0155

that the near point tests have a much more restricted range than do the far point tests. This is also graphically illustrated in Figure 2. The significance of these shifts are shown on Table 17. Although not as many of the differences in standard deviations have significance as did the differences in means, there are numerous highly significant differ-

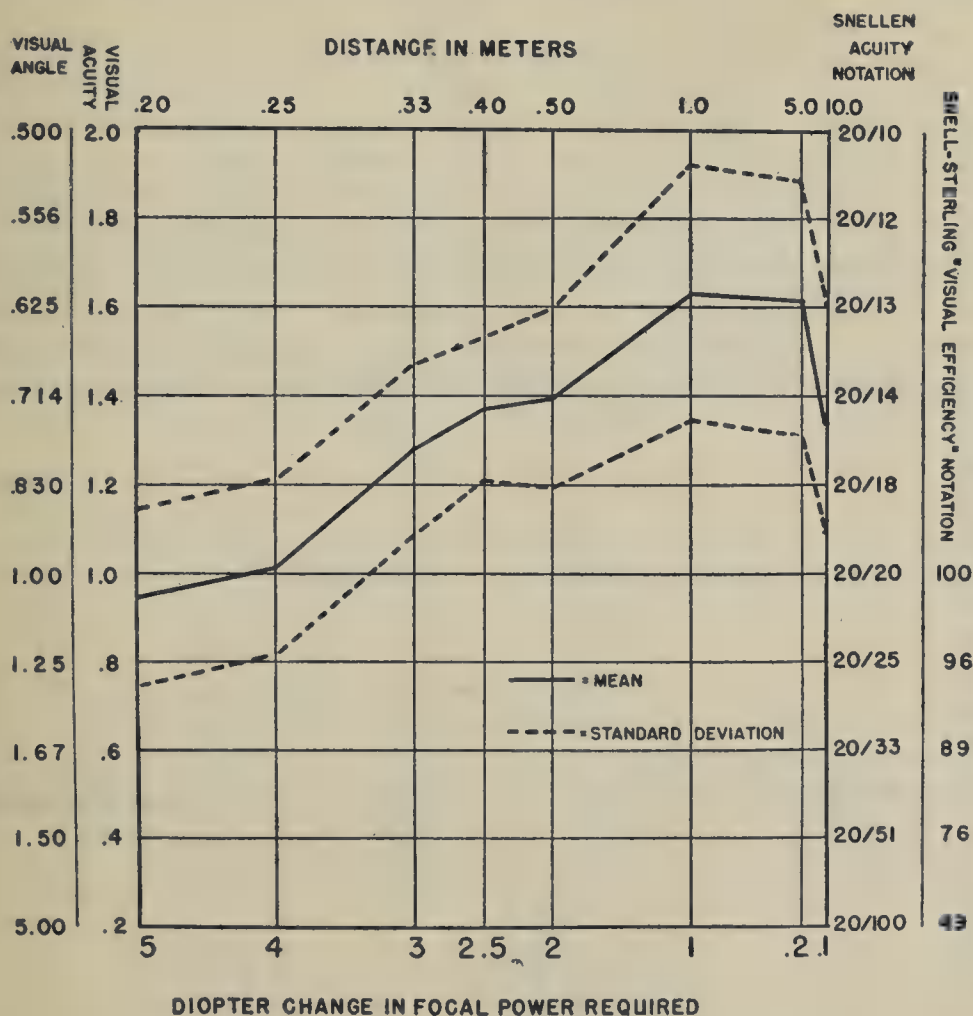


FIG. 2. Visual acuity means and standard deviations obtained on the multiple choice checker test.

ences. In general, the differences between the tests at .50 meters or less and the tests at 1.00 meters or more are very reliable.

Not only do the means and standard deviations show a consistent trend with changes in test distance but also the shape of the distribution of individual differences shows an interesting relationship with test distance. Figure 3 shows that for the two near point tests (.20 and .25 meters) the spreads of individual differences are fairly normal, but

Table 17

Significance of the Differences (Critical Ratios) in Table 16

Distance	.20	.25	.33	.40	.50	1.00	5.00	10.00
.20								
.25	.51							
.33	.92	.41						
.40	3.59	3.04	2.64					
.50	.10	.40	.82	3.48				
1.00	6.77	7.24	7.56	9.92	6.85			
5.00	8.24	8.69	9.00	11.28	8.32	1.60		
10.00	8.15	8.54	8.91	11.21	8.23	1.48	.13	

as the test distance is increased up to 1.00 meter the distributions become more and more negatively skewed. The distribution for the 5.00 meter distance is nearly as badly skewed as is the distribution for the 1.00 meter distance. The spread of individual differences at the 10.00 meters

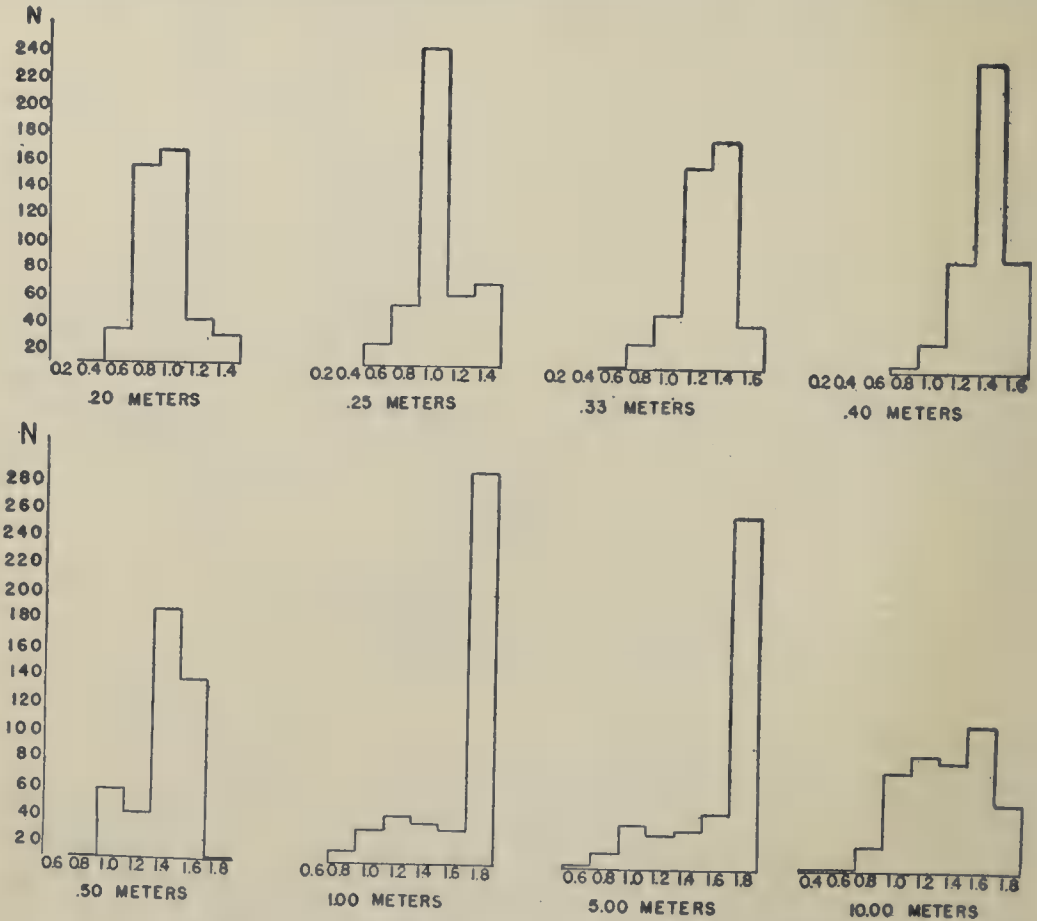


FIG. 3. Histogram on the spread of individual differences at different distances on the multiple choice checker test.

distance, although not normal, is not skewed nearly as much as the distributions for .50, 1.00, and 5.00 meters. A scanning of all of the distributions shows that for these subjects the optimum distance for acute vision is between 1.00 meter and 5.00 meters.

Conclusions and Implications

1. Although the relationship between tests of visual acuity at different distances is statistically significant, it is not high enough to presume that a measure of visual acuity at one distance will any more than very roughly predict visual acuity at another distance. This result is more clearly stated in terms of the task confronting the eye rather than in given changes in test distances in terms of feet or meters. For example, a change of only .30 of a meter from .20 meters to .50 meters is a change of 3 diopters of focal power required whereas this change at the 5.00 or 10.00 meter distance is only a fraction of a diopter, .01 and .003 respectively. The clear implication to employment office practice is that vision tests for employee allocation should be administered as close to the same distance that is required for critical vision on the job as is practically possible. In addition, the tests should not be installed on an *apriori* basis but should be validated through the same general technique used in the validation of clerical, mechanical, or intelligence tests.

2. Not only does a small difference in test distance for near vision greatly attenuate the relationship between tests but also the shorter the focal distance the lower is the general acuity. This suggests that jobs requiring constant critical near vision should be re-engineered to reduce this demand or, failing that, ocular aids should be provided since there is only a small percentage of individuals who possess the visual skills to resolve detail subtending less than a one minute visual angle.

Summary

1. The inter-relationship between visual acuity at various distances measured by the multiple choice checker design visual tests with the population of subjects used in this experiment is low, and the inter-relationship becomes steadily lower as the difference in diopters of focal power required between tests is increased.

2. This population had increasingly better vision as the focal distance was increased.

3. The spread of individual differences was more normally distributed for the near point tests than for the far point tests. The far point tests had a greater spread of individual differences expressed by either range or standard deviation.

4. The low relation between visual acuity at any given distance and visual acuity at any other distance, and the steady decrease in the relationship as the two distances are more widely separated, indicates the importance to industrial vision testing of measuring acuity at a distance at least approximately equivalent to the distance at which the employee must have satisfactory acuity. Basing the measurement of one's visual acuity entirely upon his acuity at 20 feet results in frequent errors when the job to be performed requires acuity at a much closer distance.

Received November 18, 1945.

Book Reviews

Hahn, M. E., and Brayfield, A. H. *Occupational laboratory manual*. Pp. 29. \$1.00. *Job exploration workbook*. Pp. 95. \$.96. Chicago: Science Research Associates, 1945.

Vocational counselors who have known about the courses in occupations at the University of Minnesota have been awaiting, for some time, the publication of materials which they have known must have been in the process of development there. It seemed certain that an institution doing so much in the development of tests and in the preparation of texts on counseling techniques would also make available the materials developed in its occupations courses. Hahn and Bayfield have now done this, and counselors who have looked in vain for suitable materials can now have the benefit of the Minnesota experience in the teaching of occupational information courses.

The first-named pamphlet is the counselor's or teacher's *Manual*. It discusses the plans and objectives of the Occupational Laboratory, as the course is called, the projects constituting the course, its conduct, and available counseling tools and techniques. It is emphasized that in this course the traditional recitation and examination procedure has no place, that it is, instead, a part of the counseling program and presupposes a competent counselor and extensive use of individual interviews. Assignments are tailored to the needs and interests of each individual. As the authors point out, many courses in occupations stress the presentation of occupational information: this course, on the contrary, stresses what Kitson has long urged, namely, the teaching of techniques of studying occupations and of making initial vocational adjustments. It recognizes that no student will retain the vast array of information covered by a survey of occupations. Instead, it helps him study his own problem of occupational orientation and teaches him techniques he can use again as his problems change or become clearer. The authors' suggestion that the course may, in the absence of a competent counselor, be utilized as a social science course is probably to be regretted, for too many of its values depend upon coordinated counseling.

The authors state that the materials can be used from grades 9 through 14, but only three of the 14 projects are actually recommended for use in the 9th grade. This reviewer believes that they are particularly adapted to the last year of high school and to college, and, at those levels, to students who are about to enter the world of work. There are

valuable suggestions for use in junior high school and with non-terminal students at other levels, but a number of additional projects would need to be developed for such courses.

The brief section on counseling techniques is fortunate in its discussion of counseling by fields and levels, rather than in terms of specific occupations. This fits in with the current emphasis on job families, and with the concept of vocational guidance and choice as a developmental process rather than an event. The note on personality inventories as indicators of possible difficulties in adjusting to a job rather than as guides in the choice of a vocation is in line with research on personality inventories.

But the manual is necessarily sketchy and serves as a reminder to the experienced counselor rather than as instruction of the novice. Indeed, the novice will frequently be at a loss with these materials, for example in locating job descriptions for discussion in connection with Project 5. The bibliography has other shortcomings, also, for while it includes some elementary and some advanced references for counselors, it omits other sources of equal usefulness and different emphasis and content, for example Myers' text and Clark's monograph on *Life Earnings In Selected Occupations*. Although there is some discussion of the need for briefing students before they begin field work, the counselor who has never conducted such a project with high school students may, if he does not go to more pains than suggested by the manual, be surprised at the number and types of problems that arise when inexperienced young people go out to meet the public.

The *Workbook* contains a very brief introduction for the student, and fourteen projects with explanations, directions, and forms. These projects are: vocational autobiography, former student survey, job opportunity survey, survey of employment practices, study of an occupation, investigation of training opportunities, getting along on the job, job satisfaction, job campaign, preparing a personal data sheet, filling out application blanks, evaluating employment agencies, writing a letter of application, and conducting a personal interview. Throughout, the emphasis is on laboratory procedures, implemented by such things as actual letter-writing and interviews in the last two projects, and also by a survey of job satisfaction rather than by a study of the relevant literature. In fact, better use could be made of the literature on the subject as a means of rounding out the presentation of some topics, job satisfaction among them, especially in non-urban areas with limited community resources.

In the introduction to the workbook, the fact that the students will now ask questions, rather than be asked, is probably mis-stated and over-emphasized. Actually, the student is asked a great many questions, and told to find the answers. The authors might better have stated that the

students will be asked questions, not in order to have them recite facts already passed on to them by the teacher or by a book, but rather to guide them in asking questions the answers to which will be found in real life and are important to each of them personally.

The statement that employers are often best able to help students draw their own conclusions concerning problems of vocational planning sounds too much like counseling by laymen, which is generally inadequate or misleading when not a part of professional counseling. It might better have been stated that these persons can often provide information about vocational opportunities and requirements which, when supplemented by published occupational material and by personal data obtained in questionnaires, tests, and interviews, and synthesized and interpreted by a trained counselor, will be very valuable. It is novel to find the role of test data and of the counselor insufficiently stressed in material emanating from the University of Minnesota!

A few minor suggestions might be made concerning the forms. For example, in the vocational autobiography the question, "What do I know about the occupation (of my choice)?" would be more helpful if an outline were provided of the kinds of things the student might and should know. But these deficiencies are rare. On the whole, the forms should be very useful and in need of little if any supplementation. The ability of the instructor will count most in supervising their use.

It is a pleasure to be able to say that the authors and publishers have made available two tools for the study of occupations which have been carefully prepared and thoroughly tried out, and which appear to have few if any serious shortcomings. As the authors point out, however, their usefulness will be adequately realized only when used by competent vocational counselors as part of a well-rounded vocational guidance program.

Donald E. Super

Teachers College, Columbia University

Cantor, Nathaniel. *Employee counseling*. New York: McGraw-Hill, 1945. Pp. viii + 167. \$2.00

A book on employee counseling is long overdue. There is considerable activity in this field, although the objectives and procedures of counseling in industry have remained quite vague.

Cantor urges a clear recognition that the activities of industry are social as well as economic. "An industrial organization is made up of individuals who are concerned with the economic activity of earning their livelihood, and who are also engaged in living socially."

The first part of the book states the general problem and traces the development of counseling in American industry. The attempt is made

in the second part of the book to give the reader some insight into psychological processes involved in human relationships. In the third part the author discusses the organization of the counseling staff and its relationships with employees, supervisors, the union, and management. There is a selected bibliography on industrial counseling.

The author points out that most industrial counseling programs have started since 1941. Perhaps the recency of the development accounts for some of the confusion regarding objectives and procedures. It is to be hoped that this book will clarify some of the thinking regarding the role of counseling in industry.

Industrial counselors perform one or more of these functions: Services that provide specific information to the employee; services that gather information for the personnel department; and interviews that provide employees an opportunity to express themselves. Cantor prefers to assign the counseling duties and the informational and service duties to different individuals. Such a division of duties seems desirable to the reviewer also because it clarifies the thinking of employees and counselors regarding the nature of the counseling relationship.

The discussion of the organization of personality contains such material as the following: "To be yourself, to accept your own limitations, to recognize the inconsistencies in the actions of others without feeling too hostile, and to recognize the inconsistencies in your own behavior without feeling too guilty is to approach normality of mind." According to the author, the recognition of the inherent ambivalence of every individual's behavior is the basic psychological premise underlying the counseling processes.

The counselor's sole objective is "to help the employee get rid of or lessen the intensity of the emotional burden and so free him to do a better job." In his description of the counseling process, the author reflects the influence of Rogers' nondirective therapy in such statements as ". . . the only effective way to solve an individual's problem is for the individual to face and settle his problem in his own way." Similarly, the author described the counselor as "a person who makes it possible for the employee to talk aloud, honestly, to himself."

Some of the examples of counseling presented in the book show that the counselors were amateurs and were unfamiliar with the principles outlined in this book.

With thousands of veterans being reabsorbed into industry, the need for counseling will increase. In the reviewer's mind there is still a question regarding the responsibility of industry for the psychotherapy of employees. The program outlined in this book is primarily applicable to a large industrial plant, whereas the average factory is too small to support a well-developed counseling program. It will be necessary for

community agencies to provide the professional counseling which severe cases need.

This book is valuable inasmuch as it helps clarify the objectives and procedures of employee counseling. Someone will write a much better book in a few years, however, after thinking has been clarified through more experience and reflection in this field.

Charles C. Gibbons

*W. E. Upjohn Institute
For Community Research
Kalamazoo, Michigan*

Dicks, Russell. *Pastoral work and personal counseling*. New York: MacMillan, 1944. Pp. x + 227. \$2.00.

Expanding from his previous work and writing in the field of ministerial work for the sick, the author now describes the whole of the pastoral task of protestant clergyman, with special reference to their inadequately understood counseling service. He declares the clergy is as poorly prepared for this increasingly important function as were physicians when internship began. Like physicians, counselors should strive "to cure sometimes, to relieve often, and to comfort always."

Chaplains claim that seventy five per cent of their time is given to work with individuals, their problem being to find time for the tremendous volume of possible interviews, eight to ten per day being about all a minister can handle effectively.

Pastoral calling is a fruitful procedure for initiating personal counseling, but again some successful preachers claim they are "too busy" for this. The author emphatically condemns such oversight of opportunity as a "defence for ignorance on the one hand or a lack of faith upon the other," declaring that every minister should average four calls per day.

Although declaring that advising is "dictatorship in living" the author throughout the book glibly gives considerable advice to his fellow ministers;—all of it so good one could wish it were accepted.

Part II deals in detail with the opportunities for personal counseling of the sick and dying, bereaved, unemployed, imprisoned, aged and shut-ins, new and prospective church members. "Ninety per cent of our marital counseling originates with the wife," declares the author, which may indicate the feminization of church work,—as well as the more obvious conclusion. He calls attention to the counseling axiom that no real help can be given until a counselee suffers sufficiently to want help. Pre-marital and marital counseling is rightly emphasized as appropriate work for ministers, because they perform the marriage ceremony and also because of the church's historic dicta regarding family morals. Mis-marriage is given as a prevalent cause for drunkenness.

Selecting a spouse and selecting a life-work are mentioned as the two greatest decisions that a person makes, yet very little space is given to discussing the latter. Perhaps the author's general advice applies here, "do not waste time and run risks so far as your reputation is concerned in dealing with something you know nothing about." The author's experiences in war-time counseling under the Y.M.C.A.—U.S.O., give the basis for some new emphases.

Conditions of effective pastoral work and counseling are ably considered in Part III and should give all clergymen readers clearer insight into the psychological laws governing this work. His philosophical statement that creation emanated from suffering is followed by discussion of the four types of suffering; pain, fear, guilt feelings, and loneliness. "Our reasoning powers are the first to break under the pressure of prolonged suffering," he declares and "unless viewed from the perspective of time and the larger experience of living, pain is destructive."

Two chapters are devoted to "Listening,"—presumably a difficult technique for those who by predilection and training are preachers. He points out that the confessional of the liturgical churches is limited by canon law; e.g. lying is a sin theologically but a revealing defense mechanism psychologically.

Pastors' failure to keep records is castigated as lack of requisite discipline for professional standing as counselors. Preaching is said to be rapidly declining as the principal method of carrying on the church work,—but the author outlines a splendid sermon on the counseling function and procedure of a pastor. The relation of the clergyman to other professional workers is helpfully described. While appreciating all of them, the author exalts the pastor as he whose "task is to personalize the man on the cross."

Pastors will like this helpful book, other counselors will profit by its religious approach.

J. Gustav White

California State Vocational Rehabilitation Bureau

Hudson, Holland, and Fish, Marjorie. *Occupational therapy in the treatment of the tuberculous patient*. New York: National Tuberculosis Association, 1944. Pp xii + 317. \$3.00

The authors originally planned to write a text and reference book for undergraduate students of occupational therapy. They have done that and more. They have contributed a volume which is written in a refreshing and engaging style and which should be read not only by occupational therapy students planning to work in tuberculosis hospitals, but by all practitioners and students of occupational therapy and by mature persons considering occupational therapy as a career. This book can

also be read with profit by other professional workers in tuberculosis hospitals, by vocational counselors, and by training and placement workers dealing with rehabilitation problems.

The authors, in discussing the selection of books for patients, state that "sometimes more may be learned from a text which has been prepared by a highly competent teacher than from a year's tutelage by a mediocre instructor." Certainly much can be learned from this book since the authors bring to its writing the service and experience which has placed one of them as the Director of Rehabilitation Service of the National Tuberculosis Association and the other as Director of Professional Courses in Occupational Therapy at Columbia University. Every page reflects the authors' intimate knowledge of tuberculosis, its nature, diagnosis and treatment; of the tuberculous patient and the tuberculosis hospital, as well as the role, training, and techniques of the occupational therapist in the total program. Besides giving the student a generous insight into their philosophy of therapy, the authors are realistically helpful in providing numerous practical suggestions for such specific services as library service, musical therapy, graphic and plastic arts, woodworking, household and homemaking arts, prevocational and vocational training, and placement.

Stress is placed on the treatment of the whole patient, not merely treatment of the clinical disease. The student is warned against programs in which "the patient remains an abstraction quite as if the bacillus led an existence independent of its host."

The point is also made that the treatment of tuberculosis is not a "solo performance". While the role of the occupational therapist is an important one, she cannot operate effectively except under a physician's direction and with the aid of the medical social worker, nurse and those other hospital workers whose related roles are well described. The authors deplore the fact that so few tuberculosis hospitals have the assistance of experts in mental hygiene. It is significant that, in a book which stresses the need for attention to the whole person, no mention is made of the psychologist or vocational counselor. This is due, no doubt, less to an oversight on the part of the authors than to a lack of understanding on the part of psychologists and hospital administrators of the role which the psychologist might play in this special type of dynamic and individual case treatment. If psychologists and counselors are not available to the hospital staff, some of their functions will be performed by the occupational therapist who must act also as a medical social worker in the absence of a professionally qualified person in that field. A major responsibility of the occupational therapist, say the authors, is first to study the patient and then to develop a project to fit the patient whom

she has studied, and it is in the development of practice based on this sound philosophy that this stimulating volume makes its contribution.

Gwendolen G. Schneider

*Veterans Administration,
Advisement and Guidance Service,
Washington, D. C.*

Bills, Arthur Gilbert. *The psychology of efficiency. A discussion of the hygiene of mental work.* New York: Harper & Bros., 1943. Pp. xiv + 361. \$2.75.

As the author points out in his preface, "Books on how to avoid worry and overcome fear, how to prevent and remove emotional conflicts and maladjustments, are plentiful, as are treatises on how to avoid friction in dealing with others and how to improve our personalities . . . Yet an almost complete dearth of books exists on the subject of mental efficiency; i.e., how the average, normal, well-adjusted person, geared to a daily program of work, can manage to get the most efficient service from his own mental equipment." To supply this lack, the author here brings together the principal results of a wide variety of experimental studies on mental and motor efficiency. In contrast to several books written for industrial engineers and supervisors, the author's point of view is stated as "that of the mental worker himself who would like to know how to accomplish the most with the least wear and tear and the greatest long-time satisfaction to himself."

The introductory chapter, entitled "The Thinking Machine," outlines the principle that the entire organism, not isolated segments, performs every piece of mental and physical work; "At-tention is in large part bodily tension." It gives also a preview of those aspects of work-hygiene discussed in the following chapters; and a brief resumé of these topics will serve to make clear the general nature of the book: Controlling the energy level; Mental work, fatigue, rest, recovery, 5 chapters; (fatigue is defined as any reduction of efficiency); Sleep, its nature and control; Factors in the work setting, attention and distractions, motives and incentives, emotions, suggestion, physical conditions, 5 chapters; Modification of efficiency by learning, age changes, personal organization and planning, effective thinking, 4 chapters.

The level of treatment falls somewhere between that of popular presentation and that of detailed and critical exposition of research. More space is devoted to results than to apparatus and methods, although there is enough about the latter to make the research findings intelligible, if one seeks only to understand and not to repeat or criticize the studies described. The fact that the discussions are not crowded with citations of sources and that the 168 references mentioned represent only a minor

fraction of the work done, should not bother those readers who are themselves oriented to the experimental literature, and who bear in mind the author's purposes; and it certainly makes for smoother and easier reading.

The inclusion of supplementary reading references, of 17 pages of "Test Items for Review" (true-false, completion, best-answer), and of a 15-page glossary, suggests that the author had primarily in mind classroom readers, and only secondarily those engaged in industrial supervision and planning; although for these latter also it should prove a useful guide, since practical implications of the principles have been stressed throughout. The book is written in a clear, direct, and readable style, with frequent summaries. At its level of presentation it should prove a valuable addition to the libraries of both students and professional workers in this important but hitherto neglected field.

Forrest A. Kingsbury

The University of Chicago

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Practical psychology.* Karl S. Bernhardt. New York: McGraw-Hill Book Co., Inc., 1945. Pp. 319.
- The psychology of seeing.* Herman F. Brandt. New York: The Philosophical Library, 1945. Pp. 240. \$3.75.
- Psychology for nurses.* Bess V. Cunningham. New York: D. Appleton-Century Co., Inc., 1946. Pp. 364. \$3.00.
- Reading difficulty and personality organization.* Edith Gann. Morning-side Heights, N. Y.: King's Crown Press, 1945. Pp. 149. \$2.00.
- When life gets hard.* James Gordon Gilkey. New York: The Macmillan Co., 1945. Pp. 138. \$1.50.
- Psychology of religion.* Paul E. Johnson. New York: Abingdon-Cokesbury Press, 1945. Pp. 288. \$2.00.
- Job analysis for retail stores.* M. J. Jucius, H. H. Maynard, and C. L. Shartle. Columbus: Bureau of Business Research, Ohio State University, 1945. Research Monograph No. 37. Pp. 65. \$2.00.
- New directions in psychology.* Samuel Lowy. New York: Emerson Books, Inc., 1945. Pp. 194. \$3.00.
- Prediction of the adjustment and academic performance of college students by a modification of the Rorschach method.* Ruth Learned Munroe. Stanford University: Stanford University Press, 1945. Pp. 104. Paper \$1.25, cloth \$2.00. Applied Psychology Monograph No. 7.
- Human nature and enduring peace.* Gardner Murphy (Ed.). Boston: Houghton Mifflin Co., 1945. Pp. 475. \$3.50.
- Public opinion measurement. A survey.* Laszlo Radvanyi. Instituto Cientifico De La Opinion Publica Mexicana, Donato Guerra 1, Desp. 207, Mexico, 1945. Pp. 88. \$1.00. (students \$.50)
- Measurement in today's schools.* C. C. Ross. New York 11: Prentice-Hall, Inc., 1945. Pp. 597. \$3.25.
- Guide to guidance.* Charles M. Smith and Mary M. Roos. New York 11: Prentice-Hall, Inc., 1945. Pp. 440. \$3.00.
- Where do people take their troubles?* Lee R. Steiner. Boston: Houghton Mifflin Co., 1945. Pp. xiii + 265. \$3.00.
- Mental examiners' handbook.* Revised edition. F. L. Wells and Jurgen Ruesch. New York 18: The Psychological Corporation, 1945. Pp. 211. \$4.50.

HAVE YOU RECEIVED YOUR FREE COPIES OF EDUCATIONAL BULLETINS?

- | | |
|---|--|
| No. 1. How Tests Can Improve Your Schools | No. 9. Identifying the Difficulties in Learning Arithmetic |
| No. 2. How to Select Tests | No. 10. Diagnosis in the Reading Program |
| No. 3. How to Conduct a Survey | No. 11. Appraising Personality and Social Adjustment |
| No. 4. Administrative Use of Survey Results | No. 12. Use of Tests and Inventories in Vocational Guidance and Rehabilitation |
| No. 5. Teacher Use of Test Results | No. 13. Use of Standardized Tests in Correctional Institutions |
| No. 6. Basic Testing Program | No. 14. The Proper Use of Intelligence Tests |
| No. 7. Conducting High School Guidance Programs | No. 15. Vocational Guidance for Junior and Senior High School Students |
| No. 8. Planning the Elementary School Testing Program | |

(The Above Bulletins Are Furnished Free of Charge Upon Request)

EDUCATIONAL REPORTS

- | | |
|---|--|
| Report A. The Three-R's Save a School System | Report C. Teachers and Students Improve Their Mental Health |
| Report B. A New Type Mental Test Solves Persistent Educational Problems | Report D. Arithmetic Fundamentals Test Results in High Schools |

(The Above Reports Are Furnished Free of Charge Upon Request)

Write for descriptive catalog of standardized diagnostic tests

CALIFORNIA TEST BUREAU
5916 HOLLYWOOD BOULEVARD
LOS ANGELES 28, CALIF.

Journal of Applied Psychology

EDITED BY: DONALD G. PATERSON, UNIVERSITY OF MINNESOTA

Consulting Editors

L. S. ACHILLES, *Psychological Corporation*; WALTER V. BINGHAM, *A.G.O., War Department*; AROLD E. BURTT, *Ohio State University*; ARTHUR I. GATES, *T. C. Columbia University*; HEN G. JENKINS, *University of Maryland*; IRVING LORGE, *T. C. Columbia University*; JINN MCNEMAR, *Stanford University*; WILLARD C. OLSON, *University of Michigan*; MES P. PORTER, *Swarthmore, Pennsylvania*; EDWARD K. STRONG, JR., *Stanford University*; MORRIS S. VITELES, *University of Pennsylvania*; JOSEPH ZUBIN, *N. Y. Psychiatric Institute*.

Table of Contents

<i>Studies in Job Evaluation. 3. An Analysis of Point Ratings for Salary Paid</i>	
<i>Jobs in an Industrial Plant: C. H. LAWSHE, JR., AND A. A. MALESKI</i>	117
<i>Preliminary Report on the Miami-Oxford Curve-Block Series:</i>	
P. L. MELLEBRUCH	129
<i>Comparative Study of Three Tests for Color Vision: H. FOSTER</i>	135
<i>Studies in the Application of Motor Skills Techniques to the Vocational Adjustment of the Blind: M. K. BAUMAN</i>	144
<i>The Significance of Verbal Aptitude in the Type of Occupation Pursued by Illiterates: W. D. ALTUS AND C. A. MAHLER</i>	155
<i>Readability of Newspaper Headlines Printed in Capitals and in Lower Case:</i>	
D. G. PATERSON AND M. A. TINKER	161
<i>Explorations in Personality by the Sentence Completion Method: A. R. ROHDE</i>	169
<i>Speech Intelligibility Under Various Degrees of Anoxia:</i>	
G. M. SMITH AND C. P. SEITZ	182
<i>Book Reviews</i>	192
<i>New Books, Monographs, and Pamphlets</i>	196

Published Bi-monthly by The American Psychological Association, Inc.

1500 Independence and Lemon Sts., Lancaster, Pa., and 1227 Nineteenth St., NW, Washington 6, D. C.

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the Act of March 3, 1879

Copyright, 1946, by The American Psychological Association, Inc.

Journal of Applied Psychology

Vol. 30, No. 2

April, 1946

Studies in Job Evaluation. 3. An Analysis of Point Ratings for Salary Paid Jobs in an Industrial Plant

C. H. Lawshe, Jr., and A. A. Maleski

Division of Education and Applied Psychology, Purdue University

The growing importance of job evaluation, the various approaches to the problem, and the increasing concern of the applied psychologist with this area of industrial activity have been discussed in earlier studies¹ in this series. These studies have dealt with point rating systems as they operate with hourly paid jobs, while the present one deals with a modified system used in the evaluation of salary paid jobs. The findings and interpretations in this study apply specifically to a salary rating plan in a particular plant; however, certain general principles can be found which may be inferred in other industrial situations.

Purpose of the Study

The principal purpose of this study is two-fold: (1) to identify through statistical techniques the primary factors operating in a salary rating plan as it is functioning in an industrial plant, to discover those items which tend to gravitate or cluster around the factors defined, and to determine the significance of each factor in the total point rating; (2) to determine the adequacy of an abbreviated point rating scale based on a few items, and to disclose and study any differences between the two scales in terms of their practical significance.

Rating Scale and Experimental Procedure

Salary Rating Plan. To enable its member companies to evaluate salaried positions in the same objective way that factory jobs can be rated, the National Metal Trades Association devised a salary rating

¹ C. H. Lawshe, Jr., and G. A. Satter. Studies in job evaluation. 1. Factor analyses of point ratings for hourly-paid jobs in three industrial plants. *Journal of Applied Psychology*, Vol. 28, No. 3, pp. 189-198, June, 1944. C. H. Lawshe, Jr. Studies in job evaluation. 2. The adequacy of abbreviated point ratings for hourly paid jobs in three industrial plants. *Journal of Applied Psychology*, Vol. 29, No. 3, pp. 177-184, 1945.

plan which is similar in its structure and method to the N. E. M. A. system of point rating as reported by Kress.² This plan provides for the rating of salary paid jobs on the following items: Education; Experience; Complexity of Duties; Supervision Received; Working Conditions; Errors; Contacts with Others; Confidential Data; and Mental or Visual Demand. The plan also includes two items which are used only when rating positions where supervisory duties are involved. These are "Scope of Supervision" and "Character of Supervision." This means that all other jobs receive a "zero" on the two items; actually, all jobs are rated on these items just as they are on the others.

Each item is graded into 5, 6, or 7 degrees and each degree carries a different numerical or point value. For example, the item "Scope of Supervision" appraises the magnitude of the supervisory responsibility expressed in terms of the number of persons supervised and is graded into six categories as follows:

<i>Degree</i>	<i>Description</i>	<i>Points</i>
1	Assist and direct up to 5 persons usually in the same occupation.	5
2	Supervise a small group or unit, seldom over 25 persons.	10
3	Supervise a section or department, seldom over 50 persons.	20
4	Supervise a department, from 50 to seldom over 100 persons.	40
5	Direct and supervise one or more departments, usually from 100 to 250 persons.	60
6	Direct and supervise the operations of a major division, usually more than 250 persons.	80

Points awarded to a salaried position on the item, "Scope of Supervision," are added to the values derived from the other ten items on the same job to give a total point rating for that particular job. Thus by rating each position in a plant or organization on the eleven item scale, the relative worth of each job can be expressed quantitatively in terms of total points. These totals are then converted into money values to establish a salary structure for the organization.

Source of Data. Point rating data were obtained on about 400 different salaried jobs from an industrial plant manufacturing airframes and employing over 5,000 workers. This plant also has a point rating plan in operation for its hourly-paid employees, which has been studied by Lawshe and Satter³ as Plant B.

Procedure. The salary rating data including the ratings on each of the eleven items plus the total point ratings for each job were punched on

² A. L. Kress. How to rate jobs and men. *Factory Management*, 60-65, 1939.

³ C. H. Lawshe, Jr., and G. A. Satter. *Op. cit.*

Table 1

Intercorrelations of Point Ratings of Eleven Items and of Total Points in the Job Evaluation of Salaried Workers in an Industrial Plant

	(1) Total Points	(2) Education	(3) Experience	(4) Complexity of Duties	(5) Supervision Received	(6) Errors	(7) Contacts with Others	(8) Confidential Data	(9) Mental or Visual Demand	(10) Working Conditions	(11) Character of Supervision
(2) Education	.809										
(3) Experience	.930	.756									
(4) Complexity of Duties	.928	.788	.856								
(5) Supervision Received	.860	.642	.762	.783							
(6) Errors	.864	.677	.772	.788	.771						
(7) Contacts with Others	.856	.657	.750	.768	.754	.761					
(8) Confidential Data	.717	.617	.667	.663	.607	.591	.642				
(9) Mental or Visual Demand	-.346	-.174	-.259	-.335	-.364	-.337	-.419	-.149			
(10) Working Conditions	.113	-.144	.071	.063	.097	.096	.108	-.095	-.290		
(11) Character of Supervision	.789	.494	.658	.667	.690	.630	.641	.496	-.350	.195	
(12) Scope of Supervision	.597	.289	.465	.489	.485	.429	.410	.280	-.291	.253	.703

Table 2
Factor Loadings Before and After Rotation

Rating Scale Items	Before Rotation				After Rotation			
	k_1	k_2	k_3	h^2	k_1	k_2	k_3	h^2
(1) Total Points	.991	.126	-.095	1.007	.982	.194	.081	1.009
(2) Education	.745	.436	-.127	.761	.844	-.022	-.223	.763
(3) Experience	.893	.208	-.093	.849	.914	.122	.000	.850
(4) Complexity of Duties	.914	.178	-.056	.870	.924	.117	.051	.870
(5) Supervision Received	.870	.045	.080	.765	.841	.086	.225	.765
(6) Errors	.860	.135	.149	.780	.860	-.024	.202	.781
(7) Contacts with Others	.864	.098	.213	.802	.852	-.049	.271	.802
(8) Confidential Data	.686	.330	-.069	.584	.755	-.013	-.122	.585
(9) Mental or Visual Demand	-.423	.322	-.318	.384	-.302	-.041	-.539	.383
(10) Working Conditions	.152	-.467	.178	.273	.000	.183	.489	.273
(11) Character of Supervision	.806	-.305	-.240	.800	.672	.533	.256	.801
(12) Scope of Supervision	.613	-.439	-.298	.657	.447	.621	.270	.658

machine-sort cards. Intercorrelations between all eleven items and total points were computed and a correlation matrix was prepared (Table 1). The matrix was factor analyzed by Thurstone's centroid method and three factors (k_1 , k_2 and k_3) were extracted. The extraction process was stopped when the communalities (h^2) for "total points" approximated unity (1.007). The factors and their centroid loadings are presented in Table 2 and the factor names with item groupings in Table 3. The factor loadings were transformed by the method outlined by Peters and Van Voorhis,⁴ and transformed or rotated values are also found in Table 2.

The Wherry-Doolittle shrinkage selection method was applied to the values in the correlation matrix and the three "best" items were selected for an abbreviated rating scale.

Identification of Factors

Factor Names. Of the three primary factors identified through factor analysis, Factor I was found to be common in some degree to all of the items except "working conditions." However, eight of the items stand out as possessing particularly heavy loadings in Factor I (see Table 2). These eight items are presented in Table 3 in the order of the magnitude of their loadings which range from .924 for "complexity of duties" to .672 for "character of supervision." The items identified

Table 3

Three Factors with Rating Scale Items Arranged in Order of Magnitude of Loadings

Factor	Rating Scale Item	Loading
I. Skill Demands	(4) Complexity of Duties	.924
	(3) Experience	.914
	(6) Errors	.860
	(7) Contacts with Others	.852
	(2) Education	.844
	(5) Supervision Received	.841
	(8) Confidential Data	.755
	(11) Character of Supervision*	.672
II. Supervisory Demands	(12) Scope of Supervision	.621
	(11) Character of Supervision*	.533
III. Job Characteristics	(9) Mental or Visual Demand	-.539
	(10) Working Conditions	.489

* The item Character of Supervision (11) has a relatively heavy loading in both Factors I and II.

⁴ C. C. Peters and W. R. Van Voorhis. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill Book Co., 264-268, 1940.

under Factor I all embrace some skill requirement of the individual who is to perform the job successfully and therefore it has been named "Skill Demands."

The items "scope of supervision" and "character of supervision" are heavily weighted with Factor II (see Table 2), their respective loadings being .621 and .533. These items were found to function in jobs where supervisory duties are involved. They also represent a skill demand, but it is obviously a more specific and specialized type of skill requiring supervisory ability, so Factor II has been named "Supervisory Demands." The item "character of supervision" is significantly loaded with Factors I and II and is listed under both in Table 3, since the extent of the contribution of each seems to be approximately the same.

The largest loadings of Factor III are — .539 in the item "mental or visual demand" and .489 in the item "working conditions." This factor was named "Job Characteristics" since the two items represent certain aspects of the job with which an individual must contend, and for which he should be compensated. It is significant that the item "mental or visual demand" is negatively loaded with Factor III (see Tables 2 and 3). This negative loading although not too common in factor analysis, is quite plausible and is discussed by Peters and Van Voorhis.⁵ As shown in Table 1 "mental or visual demand" correlates negatively not only with "total points" but also with each of the other ten items. In other words, whenever a salaried position in this plant was assigned a higher point value on the item "mental or visual demand," it generally received a relatively lower point value on the other items and consequently on "total points." Therefore, this item plays an important role in Factor III but in a negative direction. It is possible, by reversing the signs for "mental or visual demand," to get a positive loading. However, this would not affect the absolute value in any way, so perhaps from the logical point of view it is more meaningful to allow this loading to remain negative.

Factor Significance. Table 2 shows that the communality (h_2) for "total points" is 1.009, which since it approximates unity, indicates that the three factors extracted account for just about all of the variability in the total point ratings. If it can be assumed that these are the best rotations, then the relative proportions that each factor contributes to the total variability may be computed from the squares of the total point loadings. Thus Factor I, which has a total point loading of .982, accounts for most of the variability. Figure 1 presents these proportions graphically. Factor I, "Skill Demands" accounts for 95.6% of the total variability. Factor II, "Supervisory Demands" 3.7% and Factor III, "Job Characteristics," only .7%.

⁵ C. C. Peters and W. R. Van Voorhis. *Ibid.*, pp. 270-272.

The results in the identification of the basic factors in this scale are essentially similar to those reported by Lawshe and Satter⁶ in their study of point ratings on hourly-paid jobs in three different industrial plants. They found the factors "Skill Demands" and "Job Characteristics" present in all three plants and that "Skill Demands" accounted for most of the variance in the total point ratings in each plant. The

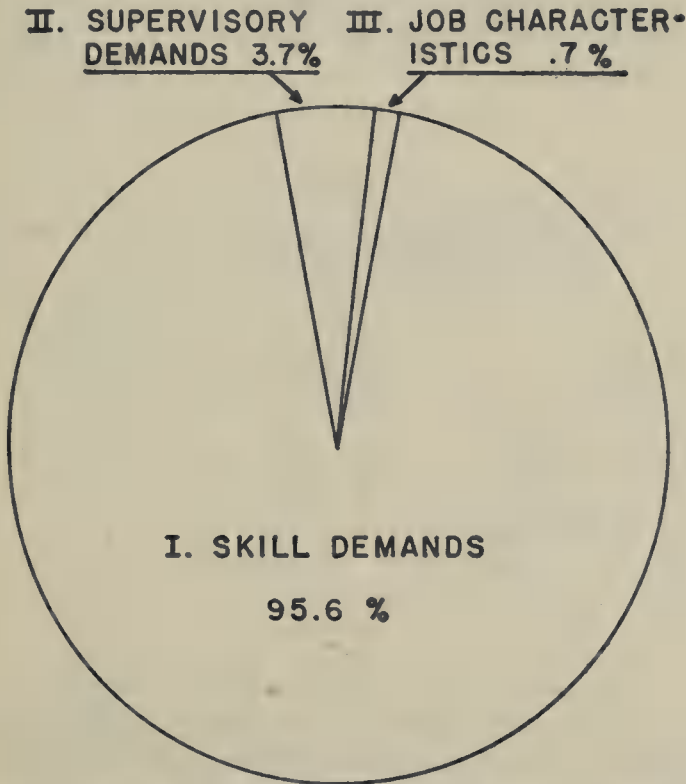


FIG. 1. The relative proportions for each of the factor contributions to the total point ratings of salary positions.

factor "Supervisory Demands" did not appear in the study. However, since it is presumably characteristic primarily of salary paid positions it would not be expected to appear in hourly-paid jobs.

Abbreviated Rating Scale

Items Selected. Of the original eleven items, the three items selected by the Wherry-Doolittle technique for an abbreviated rating scale were "experience," "complexity of duties," and "character of supervision."

They are presented with their multiple *R*'s and standard errors of estimate in Table 4. The selection process was terminated with these three items since the trial addition of a fourth variable increased the

⁶ C. H. Lawshe, Jr., and G. A. Satter. *Op. cit.*

multiple R only from .980 to .987 or less than .01. This increase does not appear significant either from a statistical point of view or in terms of sheer practicality.

Accuracy of Prediction. It can be seen from Table 4 that the variable "experience" correlates highest with "total points" having a coefficient of .930 and a standard error of estimate of 32.3 points. However, with the addition of "complexity of duties" the multiple correlation is raised to .964 and is further increased by including the item "character of supervision," to .980. At the same time, the standard error of estimate is reduced first to 26.8 and finally with the inclusion of the third item, to

Table 4

Correlation Coefficients Between Ratings on Selected Items and Total Point Ratings together with their Standard Errors of Estimate

Selected Items	Correlation Coefficients	Standard Errors of Estimate	%
Experience	.930	32.3	.32
Experience plus Complexity of Duties	.964	26.8	.26
Experience plus Complexity of Duties plus Character of Supervision	.980	20.3	.20

20.3. This means that the estimates (using the abbreviated scale) for approximately two-thirds of the jobs are within 20.3 points of the total point rating based on all eleven items. The percentage figures in Table 4 indicate the proportional size of the errors in terms of the standard deviation of the total point distribution.

Relative Contribution to Total Point Ratings. The square of the multiple R (.980²) for the three item scale is .96 which indicates that about 96% of the variance in "total points" can be attributed to the combined effect of "experience," "complexity of duties," and "character of supervision." The other eight items apparently contribute a total of 4%.

Application of the Abbreviated Scale. To determine the extent to which the abbreviated scale would give results the same as or comparable to the original scale, an analysis was made of the changes that would occur if it were actually used to rate the same positions. This was done by developing a regression equation for predicting "total points" from the three selected items as follows:

$$X_{TP} = 15 + 1.8_{Exp.} + 2.7_{CD} + 2.1_{CS}.$$

Point ratings on "experience," "complexity of duties," and "character of supervision" were substituted in the formula for each of the salaried jobs and the constant 15 added to obtain the estimated or computed total point ratings.

The computed total point values were plotted against the original total point ratings in the scattergram in Figure 2. In the salary plan

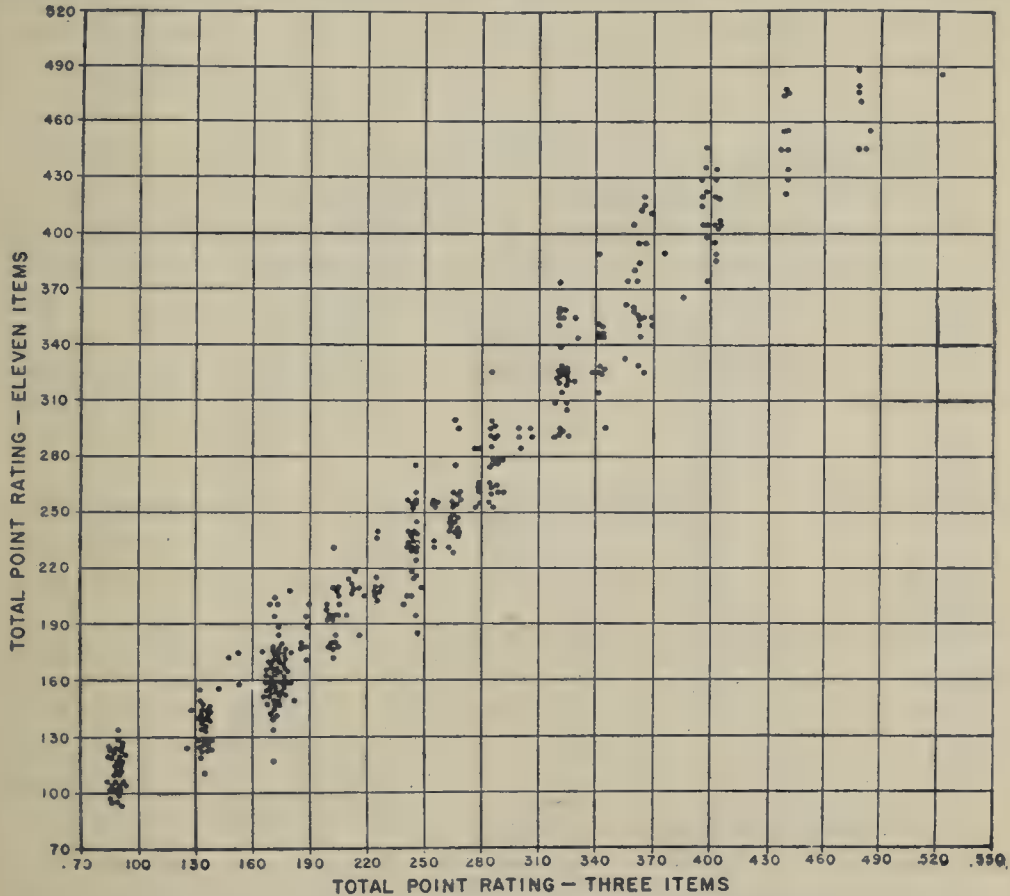


FIG. 2. Scattergram showing ratings computed from the three item scale plotted against total point ratings (all 11 items) for 391 salaried positions.

used in this plant, the rated jobs are grouped into labor grades, each having an interval of 30 points. These labor grades are indicated in the scattergram by the grid lines. Thus all jobs which were rated under 100 points fall in labor grade 1, those which received between 100 and 130 points in grade 2, and so.

Labor Grade Displacement. The deviations of the computed values for the different jobs and the number of labor grades displaced are presented in Table 5. Of the jobs studied, 44% remained in the same labor grade, 53% were displaced by one labor grade, and only 3% were dis-

placed by two labor grades. The table also shows that only 16% of all the jobs, deviate from their original rating by more than 30 points, the range of one labor grade, or the point difference between the lowest and highest rated job in any labor grade. The results obtained indicate a high degree of comparability between the two scales.

Wage Structure Considerations. An examination of the salary schedule in this plant minimizes the practical significance of the point differences that exist between the original scale and the abbreviated scale. Without reference to specific rates, for instance a certain salary may be

Table 5
Discrepancies Between Total Point Ratings (Eleven Items) and Ratings
Computed from the Three Selected Items

Points of Deviation	Percent of Jobs by Labor Grade Displacement			All Jobs
	No Displace- ment	Displaced One Labor Grade	Displaced Two Labor Grades	
0-4	16.1	1.3		17.4
5-9	13.1	8.4		21.5
10-14	8.4	9.2		17.6
15-19	4.6	6.9		11.5
20-24	1.5	7.5		19.0
25-29	0.3	6.6		6.9
30-34		6.7		6.7
35-39		3.8	0.5	4.3
40-44		1.5	0.3	1.8
45-49		0.5	1.3	1.8
50-54		0.5	0.5	1.0
55-59			0.5	0.5
Totals	44.0	52.9	3.1	100.0

paid at one time or another to employees on jobs ranging through several different labor grades. This fact tends to reduce the importance of a few points difference between the two scales.

The Reliability of Point Rating. The reliability of the original ratings is not known. It would be impractical to obtain re-ratings by a group of raters of equal ability and familiarity with the jobs. Therefore it is impossible to determine whether or not the correlation between the original and the abbreviated scales is as high as the reliability of the scale itself. Lawshe ⁷ provides some enlightening data on the reliability of point ratings in an industrial plant. The same jury of supervisors and analysts rated a group of five jobs at three different intervals within a period of

⁷ C. H. Lawshe, Jr. *Op. cit.*

about three weeks. Their ratings fluctuated anywhere from 25 to 100 points, the average change being 71 points. The relative unreliability of human judgment further diminishes the practical importance of the differences between the two scales. The results yielded by this study reveal a high degree of similarity and agreement between the original and the abbreviated scales; therefore for all practical purposes, the two scales can be considered almost identical in operation.

Summary and Conclusions

Salary rating data from an industrial plant were obtained and inter-correlations between the point ratings on each of the eleven items and "total points" were computed. The data were then factor analyzed by Thurstone's centroid method and the values were transformed or rotated by a procedure outlined by Peters and Van Voorhis. In setting up an abbreviated scale, the Wherry-Doolittle shrinkage selection method was applied. Three items yielding the highest R with the "total point" rating from eleven items were identified.

The following conclusions are supported:

1. Three primary factors were found to be operating in the rating of salaried positions in the plant and they accounted for practically all of the variability in total point ratings.
2. The factor which contributed most (95.6%) of the variance in "total points" was named "Skill Demands" since it represents skill requirements demanded of an individual doing the work.
3. A second factor, contributing 3.7% was identified as "Supervisory Demands" and includes a specific and specialized type of skill involving supervisory ability.
4. The third factor, contributing .7%, was named "Job Characteristics" because it embraces certain aspects of the job over which the employee has no control but which affect his mental or physical well being.
5. The factors identified are similar to those reported by Lawshe and Satter for hourly-paid jobs in three industrial plants. "Supervisory Demands," however, did not appear in that study since this factor is less characteristic of hourly paid jobs.
6. The items selected for an abbreviated scale were "experience," "complexity of duties," and "character of supervision." The multiple R between the three items and "total points" was .980 and the standard error of estimate 20.3.
7. About 96% of the variance in "total points" may be attributed to the three selected items; the other eight items in the scale contributed the remaining 4%.

8. If the abbreviated scale were used in this plant, 44% of the salaried positions would remain in the same labor grade, 53% would be displaced one labor grade, and 3% would be displaced two labor grades.

9. The practical significance of the differences which exist between the point ratings on the original scale and those on the abbreviated scale are minimized by the flexibility of the plant salary scale and the probable unreliability of the ratings.

10. An abbreviated scale consisting of about three items would probably give results practically identical to those obtained by the longer and more complicated eleven item scale, and would greatly reduce the time required to do the rating.

Received April 19, 1945.

A Preliminary Report on the Miami-Oxford Curve-Block Series *

P. L. Mellenbruch

Department of Psychology, University of Kentucky

The Miami-Oxford Curve-Block Series is the outgrowth of a number of years of practical experience during which the writer served as the director of a psychological clinic and was engaged in personnel work in industry.¹ Out of this experience there appeared to be need for a manipulative test involving form perception which would reveal a person's ability and his persistence in solving complex spatial problems. In the early days of his clinical work the writer designed a single curve-block which proved helpful. It is somewhat similar to the O'Connor Wiggly Block² which appeared later. It was felt, however, that a series of such blocks arranged in steps of increasing difficulty would be of much more value.

In the production of the curve-block series the following principles were kept in mind: (1) The blocks in the series should range from very simple to quite difficult; (2) the steps of increasing difficulty from one block to the next should be approximately equal; (3) the total series should not be too cumbersome but should be large enough for the smallest parts to be readily perceived and easily handled; (4) the series should be so devised as to be practically self-administering; and (5) the test should reduce to a minimum the factor of chance success or failure.

The Miami-Oxford Curve-Block Series consists of six rectangular blocks of the same over-all dimensions. Each block is cut lengthwise along curved lines so as to form a number of irregular sections. All of

* Acknowledgment is hereby given to Dr. E. F. Patten, Chairman of the Department of Psychology, Miami University, for his encouragement and helpfulness in developing the Curve-Block Series and the data here presented. The following students of Miami University gave generously of their time in the construction of the blocks and the collection of the data for this report: Robert Gehlker, Dorothy Hoffmeister, Bettie Perkins, and Robert Dixon.

¹ In 1926 the writer devised and began using the curve-block, shown in the foreground of Figure 1, in his testing of court cases and social service agency cases. This original block, somewhat smaller in size than the blocks of the present series, is cut lengthwise so as to form fifteen irregular sections. It is intended for adults and older youths and was used to help estimate the manner in which they would attack a rather difficult concrete problem as well as to indicate their persistence.

² O'Connor Wiggly Block, Johnson O'Connor (1928); Stevens Institute.

the parts of these blocks are so designed that no inside surface is flat or straight. In no case are the sections in any one block identical or reversible, nor are the sections of different blocks interchangeable.

The succeeding blocks in the series were made successively more difficult by increasing the number of cuts, hence the number of parts, in each succeeding block and also by modifying the shape of the sections in the various blocks of the series. The extent to which this has achieved the desired results can be seen by reference to the data below.

The six curve-blocks are each of a different color and are referred to in this report according to color. The *blue* block consists of three parts,

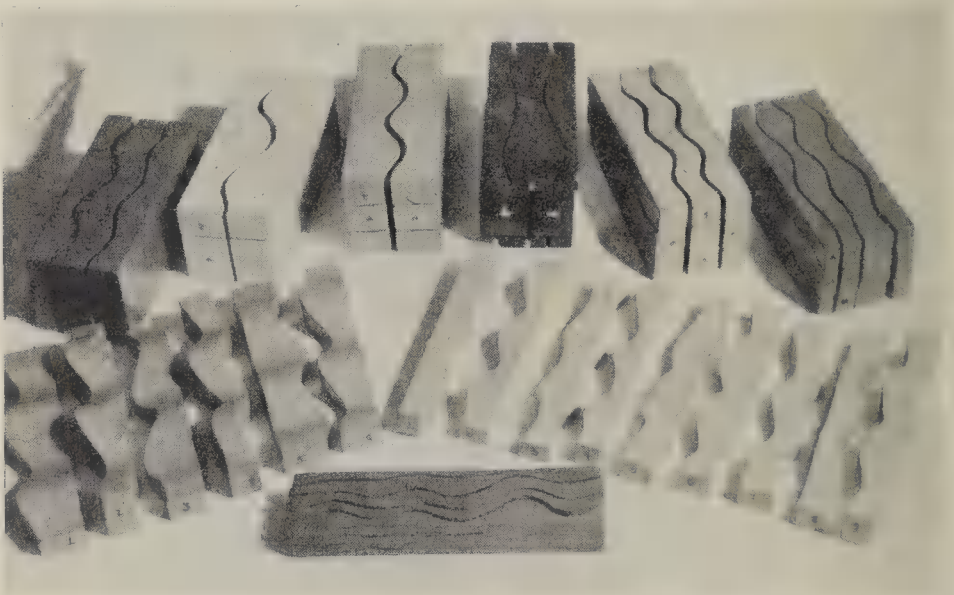


FIG. 1. Miami-Oxford Curve-Block Series (with original block in foreground).
Two blocks shown disassembled.

the *white* of four parts, the *green* of six parts, the *black* of eight parts, the *orange* of nine parts, and the *red* of twelve parts (see Fig. 1).

As a result of experimental administration of the series and the analysis of preliminary data, a number of changes were made in the size of the blocks and in the shape of their parts. The original blocks were three and a half inches square by nine inches long. These were found to be unnecessarily large so the present size was adopted, the blocks now being approximately three inches by three inches, and eight inches long. Some changes also in the shape of certain parts were made to alter their relative difficulty and to prevent sections, once assembled, from falling apart or from presenting cues for ready solution. The present series appears to be quite free from these objections.

Administration

The directions and procedures for administration are simple and are followed uniformly. Each block in the series is presented one at a time to the examinee. As the first assembled block is presented, the examiner says: "I am going to take this block apart and when I say 'Go,' I want you to put it together as quickly as you can." Then at an easy comfortable rate, the examiner disassembles the block, putting the pieces in a fixed order in front of the examinee. This is readily accomplished because the parts of each block are numbered according to this sequence.

When the examinee has reassembled the first block, each subsequent block is presented already assembled and then taken apart with the sections placed in the prescribed order. The instructions for the blocks after the first one are: "Put these pieces together as fast as you can." Such simple directions prove to be quite adequate in view of the fact that the examinee is given the easiest block first.

In this preliminary study, both the time required and the number of errors made in putting the sections of each block together were recorded. The time in each instance was accurately obtained by means of a stop watch. An error was counted if any two blocks which did not belong together, or which did not belong together in the manner attempted, were actually brought into contact with one another. Each new contact in putting the blocks together incorrectly was counted as an error even though the same error had been made before. Subsequent findings may show that the "time" score is adequate for most purposes, in which case no record need be kept of the "errors" made.

Results

It should be kept in mind that the data presented here are given in the nature of a preliminary report. The results, however, show trends which should for the most part, be borne out by subsequent findings. Further validation is planned, particularly with reference to the use of this Curve-Block Series in industry.

The data in Tables 1, 2, 3, and 4 were obtained from 93 cases, using the Miami-Oxford Curve-Block Series in its final revised form. From Table 1, it will be seen that the mean time increases from one block to the next. The proportional increases shown in Table 2 were obtained by dividing the mean time for each block by the mean time of the next easier block. This indicates, for example, that on the basis of present data the white block is 2.63 times as difficult as is the blue block; that the green block is 1.86 times as difficult as is the white block, and so forth. Stated somewhat differently we read from the second column in Table 2 that the blue block is 38% as difficult as the white block, that the white block is

55% as difficult as the green block, and so forth. These percentages of increase are obtained by dividing the mean time of each block by the mean time for the next more difficult block in the series.

While the steps of increasing difficulty are not as equal as desired, yet they are reasonably regular and may possibly prove to be more so from

Table 1
Means and Sigmas of Time Scores (Time in seconds). $N = 93$

Block	Score			Mean	σ_M	σ	σ_σ
	High	Low	Range				
Blue	2	74	72	16.0	1.30	12.56	.92
White	5	237	232	42.2	3.99	38.49	2.82
Green	12	470	458	77.5	6.99	67.43	4.94
Black	25	1018	993	170.6	16.26	156.81	11.46
Orange	71	885	814	269.3	19.23	185.45	13.60
Red	52	1730	1678	395.3	26.67	257.10	18.85
Total*	312	2735	2423	969.7	52.29	504.25	36.96

* "Total" represents the individual's score on the entire series of blocks.

further data. That these differences between the succeeding blocks of the series are significant (in terms of the "time" scores) is quite evident from the critical ratios³ given in Table 2. For the most part they are well above the requirements of real differences.

Table 2
Relative Difficulty of Blocks on Basis of Time (Time in seconds). $N = 93$

	Proportional Increase	Percentage of Increase	Critical Ratio
Blue-white	2.63	.38	6.25
White-green	1.86	.55	4.35
Green-black	2.20	.45	5.26
Black-orange	1.52	.64	3.92
Orange-red	1.47	.68	6.55

Table 3 presents the data from these same 93 cases in terms of *errors*. It will be seen that in errors as well as in time the blocks are graded in difficulty. The extent of such differences in difficulty is presented in Table 4, where it is indicated that the blocks increase rather regularly in difficulty up to the orange and red blocks, at which point the difference is not so great. And, referring to Table 2, it is seen that on the basis of

³ The critical ratio here referred to represents the difference between the means of any two blocks divided by the square root of the sum of the squares of the standard error of each mean.

mean time there is also less difference between the orange block and the red block than between other blocks of the series. The critical ratios in Table 4 show that the succeeding pairs of blocks in the series differ signifi-

Table 3

Means and Sigmas of Error Scores. $N = 93$

Block	Score			Mean	σ_M	σ	σ_e
	Low	High	Range				
Blue	0	14	14	3.7	.35	3.34	.24
White	0	22	22	6.1	.46	4.44	.33
Green	0	60	60	10.7	.98	9.42	.69
Black	0	75	75	17.6	1.79	17.23	1.26
Orange	2	142	140	32.0	2.67	25.73	1.88
Red	1	215	214	40.3	3.82	36.80	2.69
Total*	8	451	443	109.7	7.98	76.98	5.72

* "Total" represents any individual's score on the entire series of blocks.

cantly in terms of errors except for the orange and red blocks. Here there are only 96 chances in a hundred that the differences between these two means will hold. From Table 2, however, it is evident that these two blocks present significant differences on the basis of time.

Table 4

Relative Difficulty of Blocks on Basis of Errors. $N = 93$

	Proportional Increase	Percentage of Increase	Critical Ratio
Blue-white	1.68	.60	4.28
White-green	1.74	.57	4.21
Green-black	1.64	.60	3.38
Black-orange	1.82	.55	4.50
Orange-red	1.26	.79	1.77

Correlations within the curve-block series itself show the following relationships between *time* and *errors* for the different blocks and for the series as a whole:

It will be seen from Table 5 that the correlation between time and errors is rather high on four of the six blocks. This probably means that

Table 5

Coefficient of Correlation between Time and Errors. $N = 93$

	Blue	White	Green	Black	Orange	Red	Total
Correlation	.74	.72	.63	.77	.27	.54	.64
P.E.	.032	.030	.042	.028	.065	.052	.041

a time score on these four easiest blocks of the series is sufficient. However, in the light of rather low correlations between time and error on the orange and red blocks it might be of advantage to record both time and errors separately for the last two blocks of the series. This would reveal those who, in a given time, make many hurried attempts at solution as opposed to those who are more deliberate in their procedures.

The intercorrelations given in Table 6, are for the most part low enough to indicate that the blocks are sufficiently dissimilar to justify their inclusion in the series. These may indicate unreliability, however, and therefore indicate the need for a series of blocks. On the whole,

Table 6
Intercorrelations*

	White	Green	Black	Orange	Red
<i>Blue</i>	.39	.11	.28	.27	.32
<i>White</i>		.12	.11	.09	.22
<i>Green</i>			.55	.34	.19
<i>Black</i>				.42	.31
<i>Orange</i>					.43

* Odd (1, 3, 5) vs. Even (2, 4, 6) $r = .59 \pm .045$; corrected Spearman-Brown $r = .75$.

there appears to be more agreement between adjacent blocks in the series than between those removed farther from one another. While the odd-even correlation is not as high as might be desired, it still is significant. A further check on the reliability of the series by a considerable number of re-tests is planned.

It is possible that the Miami-Oxford Curve-Block Series could be used to advantage as a self-administering test. This might be done by placing the six blocks, each one disassembled with the segments in the proper order, on a long table and telling the subject to reassemble them as quickly as possible. The total time required for the series could be used as the score. Or a time limit of ten minutes or so might be used, in which case the score would represent the number of whole blocks reassembled, plus parts correctly placed in the last block attempted but not completed.

Summary

Altogether it appears that the Miami-Oxford Curve-Block Series will meet a particular need in the field of measurements, especially in industrial and clinical situations. It further appears that this series has largely met the requirements of simple administration, convenient size, steps of approximately equal difficulty, and the control of chance factors of success and failure.

Received April 13, 1945.

A Comparative Study of Three Tests for Color Vision *

Harriet Foster

University of Minnesota

Because of the importance of color blindness tests in all types of personnel work, the question of the effectiveness and the comparability of such tests is important. To the author's knowledge, only two studies have been published comparing various tests for color blindness. R. B. Philip compared the Edridge-Green, Nagel, Holmgren wools, Philip color perception, and Ishihara tests; correlations ranged from .48 to .79 for color-blind and anomalous trichomats and from .50 to .90 for all cases of defective color vision.¹ G. J. Thomas compared the American Optical Company's Pseudo-Isochromatic Plates, Ishihara, Farnsworth-Munsell, and Color aptitude test designed by the committee of the Inter-Society Color Council. Correlations were computed for color-blind and normal subjects separately, but not for the group as a whole. The most significant of these coefficients was between normal subjects on the Ishihara and the American Optical Company tests. Correlations between the various tests ranged from .14 to .64.² They show, however, only agreement *within* the classifications of color-blind and normal.

The important problem in personnel work is whether or not the tests are in agreement in their diagnosis of color-blind and normal. This can best be revealed by scatter tables which center attention upon specific instances of agreement and disagreement.

The three tests used in the present study were: (1) Pseudo-Isochromatic Plates put out by the American Optical Company, (2) Tests for Color Blindness by Ishihara, and (3) Tests for Color Blindness by Milton B. Jensen. The Pseudo-Isochromatic Test consists of 48 plates. Each plate has either one, two, or three numbers, a letter, or a curved pathway on it in a different color or a different shade of color from the background. The figures and background are in colored dots rather than a solid color. Thirteen plates of the Ishihara Test were used. They are all numbers and quite similar to the Pseudo-Isochromatic Plates. This is a well

* This study was conducted in the Psychological Laboratory, University of Minnesota, under the direction of M. A. Tinker and D. G. Paterson.

¹ Philip, R. B. A comparison of color-blind tests. *Amer. J. Psychol.*, 1938, 51, 482-488.

² Thomas, G. J. Visual sensitivity to color: A comparative study of four tests. *Amer. J. Psychol.*, 1943, 56, 583-591.

known test and has the disadvantage that some people have been able to obtain and memorize the normal responses. The test by Jensen has four plates, all of which are made to look like faces of clocks. The subject reports how many hands he sees and what hours they point to. The brevity of this test would be its chief advantage. All three of the tests have at least one plate which the color-blind and normal people see alike. This is to detect malingerers. The other plates are supposed to get different responses from color-blind subjects than from persons with normal color vision.

Each of the three tests was given according to the standard directions accompanying each test. The Pseudo-Isochromatic Plates were held approximately three feet from the subject and about level with his eyes. The Jensen test was held about ten feet from the subject unless his vision was poor. In that case the chart was held close enough so that the subject could see lines in the center of it. The Ishihara was either held at about three feet from the subject or laid out on a table with the subject standing.

Two hundred men, ranging in age from 17 to 56, acted as subjects. Women were not used because such a low percentage of them are color-blind.

The following method of scoring was used.³ Two points were counted for each color-blind response, nothing was counted for normal responses. If a number was misread, such as calling a "3" an "8" or a "7" a "2," one point was taken off since this seems to show some deficiency in color vision. If an apparently normal subject saw the number on a plate which only color-blind persons are supposed to see, one point was counted since this too seems to show a slight deficiency. The highest score possible on the Pseudo-Isochromatic Plates was 86, on the Ishihara it was 24, and on Jensen's test 6.

Table 1 shows the arbitrary classifications made of the subjects as normal, borderline, or color-blind on the three tests. More than the usual percentage of subjects were designated as color-blind by the Pseudo-Isochromatic Plates and the Ishihara test. The reason for this is probably that the method of obtaining subjects tended to attract more color-blind subjects than would be found in a random sample. Although the percentages of normal and color-blind subjects are in good agreement for the American Optical Company test and the Ishihara plates, there is still the question of whether the same subjects are color-blind on the two tests. It is also necessary to compare the Jensen test with the other two tests.

³ This method was used by Thomas, *op. cit.*

Scatter tables were made for each pair of tests. Table 2 shows the degree of agreement between the Ishihara and the Pseudo-Isochromatic Plates. There is some disagreement on the borderline case, but none as serious as to diagnose a person as normal on one test and color-blind on the other. Two cases are borderline on the Ishihara and color-blind on the Pseudo-Isochromatic, two are borderline on both tests, and one is normal on the Ishihara and borderline on the Pseudo-Isochromatic. The other 195 cases agree in the classification of color-blind and normal. This is a 97.5% agreement.

Table 3 gives the results for the Ishihara and Jensen tests. It is quite apparent that they do not agree on deviate cases. Because of the limited range of scoring units on the Jensen test, it did not seem worthwhile to designate any cases as borderline. It was difficult even to distinguish between normal and color-blind subjects. None of the subjects diagnosed as color-blind by the Jensen test was found to be normal on the Ishihara and only two of them were borderline cases on the Ishihara. But nine cases diagnosed as normal by Jensen were color-blind on the Ishihara and two were borderline. There were 187 cases in agreement in the classification of color-blind and normal. A classification of color-blind on the Jensen test is, therefore, much more valid than one of normal if the Ishihara classification is assumed to be correct. In other words, the Jensen test results may be accepted if a person tests as color-blind, but it "misses" so many cases that its value is greatly reduced.

Table 4 shows the relationship between the Jensen and the Pseudo-Isochromatic Plates. Of those testing color-blind on the Jensen, none was found to be normal by the Pseudo-Isochromatic Plates and only one was a borderline case. But again those testing as normal on the Jensen included cases which tested as borderline or color-blind on the Pseudo-Isochromatic Plates. Ten of the Jensen normal subjects are color-blind according to the Pseudo-Isochromatic Plates and five are borderline. The remaining 184 cases agree. As in Table 3, we find the Jensen test good on the subjects testing as color-blind, but not for those diagnosed as normal.

Table 5 shows the actual scores made by the 23 deviate cases. In this way it is possible to compare the three tests simultaneously. It can be seen again that the Jensen test is not in close agreement with the other two tests on these cases. Nine cases diagnosed as color-blind by the Ishihara and Pseudo-Isochromatic Plates were classified as normal on the Jensen test. Three of them actually had perfect scores on the Jensen test. Although none was found normal on the Pseudo-Isochromatic and Ishihara and color-blind on the Jensen, there is one case diagnosed as borderline on these two tests and color-blind on Jensen. The Ishihara

Table 2
Scatter Diagram Showing Scores Made on Ishihara and Pseudo-Isochromatic Plates. $N = 200$ men

Pseudo-Isochromatic Plate Scores with Diagnoses												
Normal			Borderline			Color-blind						
0-5	6-11	12-17	18-23	24-29	30-35	36-41	42-47	48-53	54-59	60-65	66-71	N
24									1	1	3	5
22-23									2	1	1	5
20-21								1	1	1		3
18-19							1		1			2
16-17								1	1			2
14-15									1			1
12-13									1			1
10-11					1			1				2
8-9												0
6-7			1						1			2
4-5	2	1										3
2-3	16	3	1									20
0-1	139	14	1	1								155
N	157	17	3	1	1	1	1	3	9	3	4	200

Ishihara Test Scores and Diagnoses

Table 3
Correlation Between Jensen and Ishihara Tests. $N = 200$ men

Ishihara														
Normal				Borderline			Color-blind							
0-1		2-3	4-5	6-7	8-9	10-11	12-13	14-15	16-17	18-19	20-21	22-23	24	N
6													2	2
5											1		3	4
4				1					1			1		3
3						1						1		2
2	34	5	1			1	1		1			1		44
1	2	1		1				1			1	1		7
0	119	14	2						1	1	1			138
N	155	20	3	2	0	2	1	1	2	2	3	4	5	200

Table 4
Correlation Between Jensen and Pseudo-Isochromatic Plates. $N = 200$ men

Pseudo-Isochromatic Plates												
Normal				Borderline				Color-blind				
0-5	6-11	12-17		18-23	24-29	30-35	36-41	42-47	48-53	54-59	60-65	66-71
Color-blind	6										1	1
	5								1	1		2
	4								1	1		1
	3										1	
Normal	2					1						
	36	3	1						1	3		
	3									3		
	1											
Normal	0	118	14	2				1		1	1	
	N	157	17	3	1	1	1	1	3	9	3	4
							0	1				200

test seems to fall down on case 10. R. E. V. This subject was classified as color-blind by the other two tests and borderline on the Ishihara. The author can offer no explanation for this. In general, it is clear that the Ishihara and the Pseudo-Isochromatic tests are in close agreement.

In testing, the experimenter found one thing that could be improved on in the Jensen test. The third plate was failed by 40 of the normal subjects. Many of the subjects said that they were confused by it because there was one curved line and one straight line. Many failed to

Table 5

Results of Three Color Blindness Tests for the 23 Deviate Cases
(Color-blind or borderline on any one test)

Individual	Pseudo-Isochromatic Test		Ishihara Test		Jensen Test	
	Score	Diagnosis	Score	Diagnosis	Score	Diagnosis
1. K. M.	71	C. B.	22	C. B.	4	C. B.
2. R. T. M.	70	C. B.	24	C. B.	5	C. B.
3. L. D. H.	68	C. B.	24	C. B.	5	C. B.
4. D. N. L.	66	C. B.	24	C. B.	6	C. B.
5. S. H.	65	C. B.	24	C. B.	6	C. B.
6. W. L. K.	63	C. B.	22	C. B.	3	C. B.
7. L. H. M.	61	C. B.	20	C. B.	0	Norm.
8. P. F.	59	C. B.	24	C. B.	5	C. B.
9. M. P. S.	59	C. B.	22	C. B.	2	Norm.
10. R. E. V.	58	C. B.	6	Bord.	4	C. B.
11. R. W. C.	56	C. B.	21	C. B.	1	Norm.
12. C. B. H.	56	C. B.	12	C. B.	2	Norm.
13. E. L.	55	C. B.	22	C. B.	1	Norm.
14. D. M.	55	C. B.	22	C. B.	1	Norm.
15. J. A. H.	54	C. B.	17	C. B.	0	Norm.
16. S. S. S.	54	C. B.	15	C. B.	1	Norm.
17. A. J. B.	53	C. B.	21	C. B.	5	C. B.
18. J. B.	49	C. B.	16	C. B.	4	C. B.
19. M. E. E.	48	C. B.	10	Bord.	2	Norm.
20. C. N.	44	C. B.	19	C. B.	0	Norm.
21. L. O.	35	Bord.	11	Bord.	3	C. B.
22. C. P.	26	Bord.	0	Norm.	0	Norm.
23. D. P.	18	Bord.	6	Bord.	1	Norm.

report the straight line because they were confused by the curved one. Even if this difficulty were removed, there would still be disagreement on the subjects diagnosed as normal by Jensen. Apparently a 3 or 4 plate test is too unreliable for individual diagnosis.⁴

⁴ The results of this study suggest, but do not prove, that a color blindness test consisting of a small number of plates may not be valid.

If the Ishihara test is accepted as a valid test of color blindness, as is customary, the Jensen test is of doubtful value in the diagnosis of color blindness. Possibly the latter is too short or perhaps the kind of response required is not as good as that of the other two tests. It might be just as ineffective with 15 plates. The only thing against the Ishihara test is its widespread use and the possibility that coaching or memory from previous exposure may invalidate the results. The Pseudo-Isochromatic Plates take more time, but it might be possible to shorten it if the only thing desired is to distinguish between the normal and color-blind. Or it may be necessary to have a large number of plates in order to be certain to identify all color-blind cases. Further research is indicated.

Summary

1. Few studies have been published comparing tests of color blindness. Because of their widespread use, it is desirable to know how the scores on the different tests agree.
2. Two hundred men were used as subjects. Each man took the Ishihara, the Pseudo-Isochromatic, and the Jensen tests.
3. The results showed that the Ishihara and Pseudo-Isochromatic Plates were in close agreement with each other, while the Jensen showed far less agreement with either of the other tests. The Jensen test mistook nine color-blind subjects for normal. If the Ishihara test is accepted as valid, the Jensen test does not have adequate validity for individual diagnosis.

Received March 5, 1945.

Studies in the Application of Motor Skills Techniques to the Vocational Adjustment of the Blind

Mary K. Bauman

Personnel Research and Guidance, Philadelphia, Pa.

The work here presented reports the application of psychological measures to the prediction of success in industrial employment for blind persons. The Trainee Acceptance Center was established to give vocational guidance to any adult who wished it in connection with training for or placement in war work. However, when the first *blind* person applied for advice, neither the experience of staff members nor the literature of vocational guidance provided adequate techniques for determining the employability of this handicapped group. Applications of psychology to the educational and more general problems of the blind, as presented in Hayes' *Contributions to a psychology of blindness*,¹ gave basic material. Early efforts in vocational guidance are reported by the same author in a later publication.²

Our attack upon the guidance of persons with deficient vision involved the measurement of skills required and consideration of the results of this measurement in the light of the training, environmental, and personality factors in each particular case. It seemed reasonable to suppose that this procedure would also yield satisfactory results if applied to blind clients.

Plan of Research

The first step was to find tests suited to measuring the skills a blind person could use on an industrial job. The fact that the Center had been established to guide individuals seeking war work oriented our early studies with the blind purely toward industry, not toward vending stands, social work, or other occupations open to them. The second step was to establish quickly the rough standards a blind person should meet on these tests if he was to be successful in industry. This was accomplished with the cooperation of a group of blind persons already employed on production who subjected themselves to the testing. It seemed possible how-

¹ Hayes, Samuel P. *Contributions to a psychology of blindness*. New York: American Foundation for the Blind, 1941.

² Hayes, Samuel P. History and present status of aptitude tests for the blind. *Unpublished Manuscript*.

ever that the learning of blind persons, particularly in handling concrete material, was so different from the learning of seeing persons that the results of two hours of testing could not be relied upon. Therefore, with the assistance of the Pennsylvania Institute for the Instruction of the Blind, the learning curves of blind persons dealing with concrete material were studied. Finally, even if the tests did appear to establish a useful standard for industrial placement, it was possible that we were merely measuring in several ways a single ability or group of abilities; a fourth step therefore involved the intercorrelation of the results of the various tests.

Tests Adapted for Use With Blind Clients

1. Minnesota Rate of Manipulation.³
 - A. Displacing. The first part of the Rate of Manipulation test was modified. The blind subject moves the blocks around the board in a pattern rather than following the standard procedure known as "Placing." Three practice and three test trials are given.
 - B. Turning. The second part of the Rate of Manipulation test was used without change except that again three practice and three test trials were given.
2. Pennsylvania Bi-Manual Worksample.³
 - A. Assembly. For the person who was totally blind or had only light perception, a preliminary trial was given in which he was asked to assemble a nut and bolt and throw them into a box just beyond the board. For persons able to distinguish the holes in the board this preliminary trial was omitted. In all cases, two trials involving placement of the assembled nuts and bolts into the holes in the board were given, and the final score used was the better of the two. On these two trials no modification was made except that the blind person was encouraged to use his dominant hand to locate the hole into which he should drop the nut and bolt.
 - B. Disassembly. This portion of the Worksample was used without modification but, again, was administered twice and the better of the two trials was used as a final score.
3. Toolsample.

This is a worksample in which various kinds of nuts and bolts are removed from one board and tightened in another with the aid of tools. Results reported in this paper were obtained by its use with male clients only; it was rarely used with female clients. Its complexity as compared with motor skills tests, and its similarity to many real jobs, make it particularly important in the evaluation of clients who have had trade experience before the loss of vision. Scoring allows for training on the first bolt if the testee cannot remove it by himself. During the remainder of the test it is possible to observe qualitatively and measure quantitatively his ability to profit by this training on the first bolt and by the experience which each succeeding bolt offers. It can therefore be used as a measure of the trainability of the individual with this type of material within the limits of a single session.

³ Descriptions of Minnesota Rate of Manipulation and Pennsylvania Bi-Manual Worksample, including norms and instructions for their use with the blind, may be obtained from the publisher, Educational Test Bureau, Minneapolis, Minn.

Tests Tried but Discarded

1. O'Connor Finger Dexterity.

This test provides useful information in the cases of persons having better than light perception but required too accurate orientation in space for our totally blind clients.

2. Purdue Pegboard.

The above statement also applies to this test.

3. Object Sorting and Assembly.

In the attempt to reproduce the fundamentals of certain jobs, a test was devised which required the sorting into separate boxes of five small objects easily distinguished by touch, and their subsequent assembly into a pattern on a stick. However, even those blind persons who performed well on other tests and who had been successful in industry could not make scores within the normal range of seeing persons on this exercise, despite additional practice. We were forced to conclude that vision gave advantages in the performance of this task which could not be overcome by any reasonable amount of practice for even the more able blind clients.

Additional Data for Vocational Guidance

A measure of general mental ability was included in every examination. For this the Wechsler-Bellevue Verbal Scale, the Interim Hayes-Binet, and the Kent E-G-Y Test were used, the test to be used in a specific case being dictated by the background and special problem of each client.

A very important part of the results in every case is the qualitative observation made by staff members during the testing. Ease in following directions, freedom from confusion, amount and kind of supervision needed are all noted. Coordination, orientation in the work space, distractibility, and motivation are also described. Staff members are encouraged to write complete descriptions of everything they observe.

A medical examination paralleling those given by industrial firms is included in the program of all clients at the Trainee Acceptance Center.

Subjects

As defined by law, blindness includes all persons having up to 20/200 vision. Our work quickly demonstrated that in the presentation of test results subdivisions of the visually handicapped would be desirable. Results of clients who are totally blind or have so little light perception as to be of no value in the test situation are presented together; this group will be referred to as "Totally Blind." Results of clients having more than light perception but not more than 10/200 vision will be presented together, and will be referred to as "10/200." Results of clients having better than 10/200 vision but not more than 20/200 will be presented together, and referred to as "20/200." In each case the notation indicates the best vision of the client, whether with one or both eyes, and with such aid as glasses could give. A division by age into a group under 40 and a group 40 and over was made because of the experience that industry favors the employment of the younger group and because all norms used at the Center are based on test results of persons under 40 years of age. Table 1 shows the distribution of subjects in each group.

Thus a total of 312 legally blind persons was studied, 231 of whom were less than 40 years of age. In this we are extremely fortunate because the preponderance of older persons in the total blind population

Table 1

Analysis of Subjects in Terms of Age and Degree of Blindness

Group	Under 40 Years of Age			40 Years of Age and Over		
	Men	Women	Total	Men	Women	Total
A. Totally blind	85	43	128	39	9	48
B. "10/200"	31	19	50	17	5	22
C. "20/200"	25	9	34	7	4	11
D. Employed group (totally blind)	18	1	19			
Total	159	72	231	63	18	81

makes it difficult even in many cities to reach a significant number of blind persons of employable age for study.

Test Results

For greater ease in handling results where sex differences are known to exist, and where raw scores can have little meaning for persons unfamiliar with the tests, all results in this section are presented in standard scores based on the testing of hundreds of seeing clients.

Tables 2, 3, and 4 give the number of persons under 40 years of age attaining each standard score, subdivided by test and amount of vision. A line is inserted between the standard score of 4.0 and scores below. This was done because this is generally considered at the Center to be the cut-off point in any trait; those with scores below this line are handicapped in any competitive situation where this trait is required. A direct positive relationship is evident between success in all six tests and amount of vision. Group C (20/200 vision) shows the largest per cent with a standard score above 4, Group B (10/200 vision) has the next largest per cent, and Group A (totally blind) has the smallest per cent. Total blindness is of the greatest disadvantage on the Displacing Test, where only 18% of the group attain scores within the normal range. It is of least disadvantage with the Turning Test where orientation in space is at a minimum.

There is in general a greater piling up of scores in the lower 2%, or standard score of 3.0 or less, than would be expected of the normal seeing population; this is particularly true with the Displacing Test. For the 10/200 and 20/200 groups, a number of individuals close to normal expectance obtain scores better than 6.0. For the totally blind group,

Table 2

Distribution of Standard Scores of Totally Blind Men and Women (Group A)

Standard Score	Dis-placing	Turn-ing	Assem-bly	Dis-assembly	Tool-sample, Part 1	Tool-sample, Part 2
7.0		1				
6.75		1				
6.5						
6.25					2	
6.0		2		1		
5.75		2		1	1	
5.5		5	1	2		
5.25		2	1	5	3	
5.0		10	3	6	1	
4.75	1	10	3	7	3	
4.5		13	3	6	1	3
4.25	3	13	6	12	7	1
4.0	2	11	16	7	6	7
3.75	2	10	13	11	5	5
3.5	6	3	10	13	9	9
3.25	5	13	11	24	12	4
3.0	4	7	18	9		8
Less than						
3.0	50	25	43	24	1	15
N	73	128	128	128	51	52

scores are concentrated below the standard score of 6.0 except for the Turning Test and Toolsample No. 1, demonstrating the importance of orientation in space on all the other tests.

Results for persons over 40 years of age, not shown in tabular form, indicate, roughly, only a little better than half as much chance that the older clients will obtain acceptable scores. There is also greater concentration of results below the standard score of 6.0; not only is there less chance of an acceptable score, but very much less chance of an exceptional score. However, no age is an absolute cutoff. One of the best records was made by a man of 54.

Since differences in amount of vision cause such significant differences in test performance and our major interest is in obtaining tests for clients not aided by vision, the remainder of this study concerns only the Totally Blind group. In addition, all further figures are based on the results from persons under 40 years of age.

Table 3

Distribution of Standard Scores of Men and Women with Better than Light Perception but not More than 10/200 Vision (Group B)

Standard Score	Dis- placing	Turn- ing	Assem- bly	Dis- assembly	Tool- sample, Part 1	Tool- sample, Part 2
7.0		1				
6.75			1	1	1	
6.5		1	1			
6.25		1		2		
6.0	1	1		1		
5.75		1	1	4		
5.5		4	2	2		
5.25	2	3	1	1		
5.0			2	3		
4.75	2	7	6	3	1	1
4.5	3	4	4	5		1
4.25	6	6	4	11	3	1
4.0	5	3	4	3	2	1
3.75	3	4	7	4	2	2
3.5	6	3	6	2	2	4
3.25	4	1	4	2	3	1
3.0	4	2	1	3		
Less than 3.0	13	8	6	3		5
<i>N</i>	49	50	50	50	14	16

Calculation of correlation coefficients between *mental ability scores* and *motor skills scores* seemed unwise because three different tests had been used as mental measures, because these tests had not all been given by the same tester, and because the tests themselves are all somewhat inadequate as measures of the mental ability needed for success in industry. The totally blind group was therefore roughly separated into three sections: superior, having I.Q.'s of 110 or better; average, having I.Q.'s between 90 and 110; and inferior, having I.Q.'s below 90.

Distributions for each mental ability group on the six tests were prepared.⁴ Mental superiority seems in general to have some advantage, mental inferiority some disadvantage. However, there is little difference in spread of scores between the average and superior groups. On Disassembly, Displacing, and Toolsample No. 1, the highest scores were

⁴ All distributions mentioned in this paper are on file with the Library of the American Foundation for the Blind, New York, and can be borrowed by application to them.

Table 4

Distribution of Standard Scores of Men and Women with Better than 10/200 but not More than 20/200 Vision (Group C)

Standard Score	Dis-placing	Turn-ing	Assem-bly	Dis-assembly	Tool-sample, Part 1	Tool-sample, Part 2
6.75		1		2		
6.5		1	2	1	1	
6.25		1		2	1	
6.0	2	4	2	1	1	
5.75		1	3	4		
5.5	3	3	3	5	1	
5.25	1	2	3	5	1	1
5.0	2	3	4	4	2	1
4.75	2	5	3		1	1
4.5	4	1	1	1	2	3
4.25	3	3	2	2	1	
4.0	3	1	2	4	3	
3.75	4	3	2		1	3
3.5	1	1	5			2
3.25	3	2			1	1
3.0				1		1
Less than 3.0	5	2	2	2		3
N	33	34	34	34	16	16

obtained by persons in the average group. Also, the best score for Turning for the mentally inferior group exceeds the best score for Turning for the mentally average group, while the best scores for Assembly for both groups are at the same point. Here again, as in the case of age, generalizations on the basis of mental ability may be made for the group but could be most unjustly applied to a given individual.

To study the significance of *age of onset of blindness* the group was subdivided into five parts: those blind at birth or before reaching one year of age, those who became blind at ages between one and nine years, those who became blind between 10 and 19 years, those who became blind between 20 and 29 years, and those who lost their vision after 30. Since our clients varied in age from 15 to 39 years, this type of subdivision did not indicate the length of time each group had been blind. A second division was therefore made, comparing the test results of persons blind more than ten years with those of persons blind less than two years.

On Displacing, Turning, Assembly and Disassembly, there is a slight but persistent advantage for those blinded early in life. Except on Turning, this advantage is even clearer when those blinded more than 10 years are compared with those blinded less than two years. Not only do a greater proportion of those blinded early in life or those blinded more than ten years before testing attain acceptable scores, but most of the really high scores are obtained by these groups. Both quantitatively and qualitatively a longer period of blindness results in a distinct advantage for the group.

With the two parts of Toolsample the reverse is true. Despite the preceding evidence that early blindness or prolonged blindness leads to improved orientation and general adjustment to manual tasks, a distinct superiority on Toolsample is shown by those groups which lost their vision in their teens or later.

A tabulation of results grouped by *cause of blindness* indicated no significant differences. Apparently a client is likely to do equally well on these tests whether the diagnosis of his blindness is Retinitis Pigmentosa, trauma, or any other cause. This may be partly explained by the fact that the causes of blindness are sometimes obscure, and in many cases more than one factor plays a part.

Study of Employed Group

In order to establish roughly some standard which would enable us to make use of our test results in recommending placements, before the completion of this study, nineteen blind persons who were successfully employed in industry were asked to take the tests.

As with persons having vision who have been successful in industrial employment, the members of this employed blind group rarely obtain standard scores below 4.0. In no case did one of them obtain more than one standard score below 4.0 out of five tests. The Displacing Test which subsequently grew out of our experience in testing blind persons, had not been developed at the time this group was tested. In no case did a client in this group deviate by more than one and one-half standard score units from the mean of the seeing population. As far as the test results are concerned, this, then became our criterion—that for prediction of successful employment in industry, the blind individual should be able to attain on these tests standard scores of 4.0 or better, or should fall below that level only slightly and only on one test.

Study of the Learning Curve

To determine the curve of learning of blind persons in repetitive manipulations, 51 blind students were given a period of training with the

Turning part of the Minnesota Rate of Manipulation and the two parts of the Pennsylvania Bi-Manual Worksample. Of these students, 27 were girls, 24 boys. They ranged in age from 14 years 7 months to 27 years 6 months, with an average age of 18 years. This period of training consisted of the repetition of the three tests, three times without rest between, on each of four successive days. The training was done under controlled conditions, where no interruptions or distractions were permitted. The establishment of equal motivation was, unfortunately, more difficult. It was not possible to work out any material reward which would be acceptable under the circumstances. The potential importance of this experiment was explained to the students but it is difficult to determine the extent to which this completely abstract motivation operated with students of varying degrees of maturity, varying intellectual capacities and varying backgrounds. In many cases great interest was expressed but in a few instances the reverse was true.

Despite this unfavorable circumstance with regard to motivation, the results approximate the familiar outlines of curves of learning found in many other studies. Ten students from a public high school were given the same training under similar conditions. Their results follow very closely the curves of the blind students.

Intercorrelations of Test Results

Since the test results for blind clients under 40 years of age do not fall into a normal distribution, the rank order method was used in obtaining the intercorrelations in Table 5. Further, because of certain changes made in methods of testing, only about half of the sets of results could be used to obtain these correlations.

All the correlations, excluding Toolsample, parts 1 and 2, are higher than those found in our studies of seeing clients at the Center. No statistical evidence is available at this writing to explain this but our

Table 5

Intercorrelations of the Test Results of Totally Blind Persons on the Six Manipulative Tasks

	Turning	Assembly	Dis-assembly	Tool-sample, Part 2	Tool-sample, Part 1
Displacing	.70 (70)	.72 (70)	.74 (70)	-.08 (32)	-.42 (32)
Turning		.58 (71)	.48 (71)	-.04 (32)	-.48 (32)
Assembly			.74 (71)	.27 (55)	-.10 (55)
Disassembly				.35 (55)	.02 (55)
Toolsample, part 2					.62 (55)

Note: The number upon which each coefficient is based is shown in parentheses following it.

observation suggests that the important part played by orientation in the work space increases the coefficient of correlation. This is especially evident in the correlations of the Displacing Test, in which it has already been noted in other connections that orientation plays a vital part. The negative correlations between Displacing and Turning versus Tool-sample are probably further evidence of the relationship mentioned under the discussion of test results related to length of blindness; in our heterogeneous group, a number of the persons who were most at a disadvantage with the first two tasks had the advantage of some familiarity with the materials of the Tool-sample.

Summary

1. Among persons who are legally classed as blind, the group which retains some degree of vision has a distinct advantage in the performance of concrete tasks, chiefly because of the ease with which they orient themselves in the work space. This generalization with regard to the group, however, does not necessarily apply to the individual. Many totally blind persons surpass others having close to 20/200 vision, and the amount by which they surpass them varies from task to task. Indeed, those blind persons who obtained standard scores above 5.00 surpassed at least fifty per cent of the seeing population upon which the norms were established.

2. Being more than 40 years of age may also be regarded as disadvantageous if group results only are considered, but in certain individual cases no disadvantage is found.

3. Mental inferiority is a disadvantage but it is not yet possible to set a definite level below which no individual should be considered for job placement and above which most will be successful. This situation may be mainly the result of the inexactness of the measures of mental ability which have been used with the adult blind.

4. Studies of test results in relation to age of onset of blindness suggest that orientation in space comes slowly. They also point very strongly to the desirability of including in the education of blind children far more contact with practical materials of possible vocational value, such as tools. The inability of some blind individuals to deal with this type of material suggests lack of previous contact with it rather than lack of fundamental ability to comprehend it.

5. Our studies give no reason for basing job placement on the cause of blindness as reported by the individual.

6. The curve of learning for blind persons takes the same general course as does that of seeing persons. The blind, however, require a greater number of repetitions before reaching their maximum performance. As a group, the blind will approximate the performance of the

seeing most nearly in those motor skills in which orientation in space is of least importance, and will remain farthest from the standard of the seeing where orientation in space is most important. Individual members of the blind group may, however, give an excellent performance even where orientation in space is a major factor.

7. In general, a higher correlation must be expected between tests of motor skills used with blind persons than between these same tests when used with seeing persons. This probably is due to the fact that blind individuals with superior orientation in space have an advantage on all motor skills tasks.

8. In testing and experimentation upon motor skills, motivation is more difficult to establish with the blind than with the seeing. The blind as a group live in a protected world where speed is relatively unimportant. In some cases, the very concept of speed seems foreign to them.

9. Tests can be of great value in the selection of individuals for job placement and in the selection of the job in which a chosen individual shall be placed. Their predictive value is by no means perfect, but that is also true of their predictive value when used with seeing groups. A more comprehensive group of motor skills tests is needed. A better measure of general mental ability, particularly sensitive to comprehension rather than mere verbalization, is needed. And, perhaps most of all, a reasonably reliable measure of emotional and personal adjustment is extremely needed.

10. Two points require final emphasis. The first is that vocational testing in general, and testing of the blind in particular, is still so inexact that it can be safely used only by a person with sound psychological training and good clinical experience. The importance of the qualitative aspects of the testing was mentioned earlier. This is a factor which cannot be subjected to statistical treatment or readily reduced to a written statement. Unless the test administrator has, from his clinical background, learned to make reliable observations and judgments of people, and unless tests have a meaning for him which goes beyond the bare words of an instruction manual, he should not jeopardize the future of a blind person by a rash diagnosis.

11. The second point is also an amplification of one that is true for all vocational guidance. The individual client is the only point of consideration. Studies like the present one lead to generalizations about the group but guidance and placement work deals with individual men and women. Variations of individuals from the average of the group are tremendous. Not tests alone, but medical findings and history, social, economic, and educational background, personality factors, and indications of present motivation, are vitally important considerations.

Received February 7, 1945.

The Significance of Verbal Aptitude in the Type of Occupation Pursued by Illiterates *

Capt. William D. Altus and T/Sgt. Clarence A. Mahler

*Ninth Service Command Special Training Center, Camp McQuaide, California***

The discharged veteran is accorded, in most cases, the right to further his education for one to four years, if he so desires, with government assistance. Proper vocational and educational counseling of soldiers returning to civilian life is highly important. One type of soldier who especially needs expert guidance if he desires vocational or other training is the soldier who was classified as an illiterate when he was inducted into the Army. If this man is to be tested as part of the guidance procedure, account must be taken of his reading handicap.

The universally employed group test may have adequate validity for those who are normally literate, but for those whose literacy is marginal the group test score is not always dependable as a criterion of aptitude. All soldiers, for instance, take the Army General Classification Test. In one study of soldiers processed at this Center for Army illiterates, there was an average increase of approximately twenty points in standard score on the AGCT from their first testing, prior to entry into the Center, until they were tested as graduates some weeks later. It is reasonable to assume that if these men had had somewhat more training in this or a similar center, where twelve weeks was the maximum time they could stay, they would show an even higher average increase in AGCT score.

In order to obviate some of the gross effects of illiteracy, as reflected in an aptitude score on a group test, the Personnel Consultants' Section in the Ninth Service Command Special Training Center used the Wechsler Mental Ability Scale, Form B, an individually administered test, as the criterion of verbal aptitude. This test was supplanted by the Army Individual Test in late 1944. Not all the subtests of the Wechsler were given; in fact, if the trainee was English speaking, he was given only four of the verbal subtests, Information, Arithmetic, Comprehension and Similarities. In this shortened form the scale proved to possess adequate validity in a battery of tests for initially selecting those men who

* The opinions expressed in this article are the authors' and do not necessarily reflect the official attitude of the Army of the United States.

** Center inactivated in November, 1945.

could not qualify as literate within the twelve weeks' time limit. The coefficient of biserial correlation of the Wechsler with the disposition of the trainee (whether graduated or discharged as inapt) is .521 for over four thousand cases. It is therefore clear that this individual test discriminated among the abilities of trainees so far as their disposition is concerned, even though all of these soldiers upon arrival were technically illiterate as the Army defines illiteracy.

It should be mentioned that, although these trainees are initially classified as illiterate and are therefore presumably homogeneous in that none reads very efficiently, they are not so homogeneous in verbal aptitude (or intelligence as it is commonly called). On the contrary, this Center may be said to have received approximately the lowest three sigmas, as measured on the Wechsler test, if the whole range of verbal aptitude among those who are self-supporting is divided into six sigma units. These men showed considerable variation in mental ability, much more than might be inferred from their classification as technical illiterates. Among some of the most capable of all these men were a few who were so illiterate that they could recognize few, if any, printed letters. Even if they were eventually graduated and shipped, a group aptitude test did these latter men a rank injustice.

In addition to pointing out the inadequacy of group aptitude tests for some of these technical illiterates, it is the purpose of this paper to present data concerning the validity of the Wechsler standard scores in terms of the occupations these trainees have pursued. These data can fortunately be obtained from the personnel card which was filled out for each trainee upon his arrival. Many individually administered tests, besides the Wechsler, were given the trainee; also, certain biographical data were obtained and recorded on the personnel card. Among the biographical data were included the occupations at which the trainee had been employed prior to his induction. His main job, the one he followed longest during his work history, is the one used in the present study. Frequently, his job immediately prior to induction was a short-term one in some war industry, one to which he obviously would not return as a peacetime occupation and one for which, in many instances, he would not have been hired if labor had not been extremely scarce. To be classified as a main occupation, a job must have been pursued for a minimum of one year, even for the youngest of trainees, though usually the amount of time spent on the job was at least three to five years. Only Negroes and old-line, native-born Whites were used in the tabulation in order to avoid the effect of language handicap among certain bilinguals as reflected in their tested ability. The personnel cards of 1,674 White and 790 Negro trainees were reviewed for this study.

There were altogether 142 different job categories for the two racial

groups. The Whites had held 114 different jobs; the Negroes, 78. The list of these jobs is too long to present here but the sixteen main occupations, which account for 81.07% of all the trainees studied are presented in Table 1. The larger list of total occupations included horse trader, entertainer, casket upholsterer, pinsetter, boxer, tree surgeon, motion picture projectionist, shepherd, fur skinner, florist, dental technician, barrel maker, bookbinder and so on.

It will be seen in Table 1 that agricultural pursuits claimed over half of the 1,998 trainees here presented. There are significantly more Whites in the agricultural group than Negroes; the critical ratio of the difference in percentages is well over the conventional three. Though there were more farmers among the Whites, there were more Negroes who had engaged in truck driving. The Negroes also show a higher proportion of general and construction laborers. More Whites had been engaged as tractor drivers but there was a greater incidence, percentagewise, of cooks and mechanics' helpers among the Negroes. One occupation, that of porter, appeared to operate solely through color selection, for while 33 Negroes had been so engaged none of the Whites appeared in this category. For the other occupations listed in Table 1, the group differences in percentage are not large.

Table 1

The Sixteen Main Occupations of White and Negro Trainees

Occupation	White		Negro		Total	
	N	% of all White	N	% of all Negro	N	% of Both Negro and White
Farm Hand	856	60.75	194	32.94	1050	52.56
Truck Driver	187	13.27	115	19.52	302	15.13
Laborer, Gen'l	85	6.03	73	12.39	158	7.91
Laborer, Const.	48	3.41	33	5.60	81	4.05
Cook	18	1.28	32	5.44	50	2.50
Mechanic's Helper	28	1.99	22	3.74	50	2.50
Tractor Driver	34	2.41	5	.84	39	1.95
Miner	34	2.41	4	.68	38	1.90
Sawmill Worker	14	.99	22	3.74	36	1.80
Porter	0	.00	33	5.60	33	1.65
Section Hand, R. R.	20	1.42	12	2.04	32	1.60
Service Station						
Attendant	12	.85	20	3.39	32	1.60
Painter	18	1.28	11	1.87	29	1.45
Welder	20	1.42	7	1.19	27	1.35
Lumberjack	19	1.35	2	.34	21	1.05
Carpenter	16	1.14	4	.68	20	1.00
Totals	1409	100.00	589	100.00	1998	100.00

When the whole group of 2,476 trainees was classified by major occupational groups and divisions of the *Dictionary of Occupational Titles*, it was found that one in eleven had held a skilled job, such as carpenter, baker, miner, auto mechanic, painter, plumber, tailor and the like. One in four held a semi-skilled job, such as tire capper, roofer, presser, machinist's helper, lumberjack, truck driver or carpenter's helper. Roughly one in seven held an unskilled job. The other four broad groupings included the agricultural, fishery, forestry and kindred occupations (one broad category), the service occupations, clerical and sales and, lastly, the professional and managerial. The fact that there were 25 trainees in this last category arises from the peculiarities of the classification system: Boxers, dancers and entertainers are classified with the professional group. Two belonged legitimately to the managerial class; one of them owned and ran a cafe and another managed a lumber yard. Only six were classified with the clerical and sales group while 154 were in the service occupations. About 40% of the total fitted into the agricultural, fishery, forestry group.

Those engaged in agriculture were classified with the unskilled for the brief study which follows. The average standard score on the four verbal subtests of the Wechsler was then computed for the skilled, the semi-skilled and the unskilled, the Whites and Negroes being separately tabulated. When the tabulation was completed, it was found that the skilled and semi-skilled workers among both the Whites and Negroes had a Wechsler mean score which was significantly higher than that of the unskilled. There were also 87 chances in 100 that the skilled Negro worker's mean aptitude score was reliably higher than that of the Negro semi-skilled worker; for a similar comparison among the White groups, there were 96 chances in 100 that skilled worker's higher mean score was not due to chance factors. Even though all are technically illiterate when they arrive at the Center, it is clear that the skilled and semi-skilled workers, whether White or Negro, are reliably brighter than the unskilled if the Wechsler mean score is equated with brightness. The Wechsler also comes close to differentiating the skilled worker from the semi-skilled in terms of verbal aptitude.

The second study was addressed to the extremes in tested aptitude. The brightest ten per cent, among both White and Negro groups, earn approximately 28 or more points in standard score on the four verbal subtests of the Wechsler while the lowest ten per cent earn from zero to twelve points. There were 100 men in each of the four categories, that is, there were 100 White and 100 Negro trainees scoring from zero to twelve points and a like number scoring 28 or more points. The categorizing as skilled, semi-skilled and unskilled was accomplished, wherever possible, by the use of the *Dictionary of Occupational Titles*. Where the

men did not fit into one of three categories of skill but fitted into one of the other broad occupational groupings, they were shifted into one of the three divisions of relative skill by a procedure that could be called rule-of-thumb.

Table 2 shows that three times as many Whites and almost twice as many Negroes scoring 28 or more points on the Wechsler were classified as skilled as was true of the comparable low-scoring group. Almost three times as many of the higher-scoring Whites were classified as semi-skilled as was found among the low-scoring group; the differential was even more marked for the Negroes. Almost four-fifths of the low-scoring

Table 2
Occupational Skills and Extremes in Wechsler Aptitude

Group	Skilled		Semi-Skilled		Unskilled	
	Bright	Dull	Bright	Dull	Bright	Dull
White	18	6	45	17	37	77
Negro	9	5	49	15	42	80

White and Negro groups are to be found among the unskilled, roughly twice the figure found for the high-scoring trainees. Aptitude, even among the technically illiterate, varies significantly according to the type of occupation pursued.

Not shown in Table 2 is the number of occupations engaged in by these groups of varying aptitude. The 100 relatively more capable Negroes had been employed in 39 different occupations; the 100 relatively duller Negroes had engaged in only 30 occupations. For the White groups the figures are even more marked: The high-scoring group had engaged in a total of 37 different occupations, while the low-scoring group had engaged in only 18. This finding is consistent with what would logically be expected, that is, that the brighter the individual the greater the range of occupational choice open to him. Verbal aptitude thus appears to be associated with not only the type of occupation followed by these technically illiterate men of limited mentality but also with their range of occupational choice.

One can never be certain of interpreting correctly a low score on a group test of aptitude when the score is earned by a marginally literate individual. Literacy and verbal aptitude are positively correlated, to be sure, even in such a restricted group as is found in an Army Special Training Center. Nevertheless, the correspondence is not sufficiently high for one to assume that all illiterates are automatically dull. As has been previously said, some of the most verbally apt trainees in Wechsler score have arrived at this Center total illiterates.

Only eight to ten minutes are required to administer the shortened form of the Army Wechsler here employed. Though taking little administration time, the four subtests have been found to discriminate quite satisfactorily among the varying learning capacities of the trainees, so far as the acquisition of literacy is concerned, and also to be associated with occupational levels, the brighter in score tending much more often to have followed a more highly skilled job. For these reasons, it is recommended by the authors that those who counsel these men when they return to civilian life use an individually administered test of comparable validity as a necessary part of the evaluative process, particularly if the counselling is concerned with occupational choice or occupational training.

Received February 3, 1945.

Readability of Newspaper Headlines Printed in Capitals and in Lower Case

Donald G. Paterson and Miles A. Tinker

University of Minnesota

The question as to the relative legibility of headlines set in all capital letters and of headlines set in lower case has long been in dispute among newspaper men. In the absence of scientific studies of the problem it is obvious that the question must remain one for debate.

In the course of our investigations of typographical factors influencing speed of reading we demonstrated that paragraphs printed in all capitals are read 11.8 per cent more slowly than paragraphs printed in caps and lower case. Few typographical factors exert such a retarding effect on reading speed.

In presenting these results in *How to make type readable*,¹ the writers strongly recommended that all capitals printing be eliminated where speed of perception is at a premium. It was pointed out that printing in all capitals survives in those situations where speed of perception is especially important namely, posters, car cards, billboards, newspaper headlines, etc. The headline writer was urged to abandon all capitals and to rely primarily on lower case type.

Within a few months after those recommendations were published, a long and lively debate by correspondence between the writers and Theodore Bernstein of the *New York Times* was initiated by Bernstein who challenged our right to generalize from the normal reading of continuous text to the perception of short headlines.² He pointed out the essential differences between these two situations and argued that all capital headlines would be perceived more quickly and at a greater distance.

The debate was terminated by mutual agreement that the question should be settled by direct experimental attack.

The writers persuaded two graduate students to undertake a first study of the question by means of the orthodox brief exposure or tachis-

¹ Paterson, D. G., and Tinker, M. A. *How to make type readable*. New York: Harper and Brothers, 1940, pp. 22-25. (Obtainable from the authors.)

² Curiously enough, no mention is made regarding the relative legibility of all capitals vs. lower case headlines in the excellent book by Garat, R. E., and Bernstein, T. M., entitled: *Headlines and deadlines*. New York: Columbia University Press, 1933.

toscopic technique. Five word single column headlines set in 24 pt. Cheltenham bold face were printed in all capitals and in caps and lower case and the relative legibility of the two kinds of headlines was determined at the normal reading distance of about 15 inches. Breland and Breland's findings³ disclosed an 18.9 per cent difference in favor of lower case single-column headlines. English likewise found a similar difference.⁴ In other words, lower case printing was found to be even more efficient in the single-column headline situation than in the continuous text situation.⁵

Realizing that Breland and Breland's findings might not hold for single-column heads when read at a distance, nor for multi-column or banner headlines which are intended to be read from afar, the writers persuaded another graduate student (Miss Alice Warren) to undertake a study of these additional problems. There was no assurance that a similar result would be found. Indeed, an opposite result might well have been expected in view of Tinker's previous demonstration that "both capital letters and words in capitals were read at greater distances from the subject than letters or words in lower case."⁶

The reader should keep in mind, however, that the greater perceptibility at a distance of capital letters and words in capitals was a function of perceptibility without reference to *speed* of perception. Furthermore, the experiment was confined to the perceptibility of isolated letters and isolated words and not to the legibility or readability of meaningful phrases such as are used in multi-column headlines. The headline problem is essentially one of the apprehension *at a glance* of the meaning of a group of related words. Under these conditions single-column and multi-column headlines in caps and lower case might be found to be apprehended at a glance even at a greater distance than when such headlines are set in all capitals.

Warren's first study utilized the same single column headlines as were used in the Breland and Breland study but instead of being exposed by means of the Dodge Tachistoscope at a distance of 15 inches from the

³ Breland, K., and Breland, M. K. Legibility of newspaper headlines printed in capitals and lower case. *J. appl. Psychol.*, 1944, 28, 117-120.

⁴ English reports a reading loss of 18 per cent for headlines set in regular all caps in a comparison with headlines set in caps and lower case. See E. English, A study of the readability of four newspaper headline types. *Journalism Quarterly*, 1944, 21, 217-229.

⁵ In a letter to M. K. Breland, Bernstein wrote, "Even though I must confess that the results showed a more pronounced difference than I had expected, I am gratified that at least I had some part in advancing scientific backing for a hypothesis that had been airily affirmed for years without such backing. Your paper is important in the annals of both typography and journalism."

⁶ Tinker, M. A. The influence of form of type on the perception of words. *J. appl. Psychol.*, 1932, 16, 167-174.

subject, they were exposed by means of the Whipple Disc Tachistoscope at a distance of five and one-half feet from the subject.⁷ The exposure time in the Breland and Breland study was 50 milliseconds (1/20th of a second) whereas in the Warren study the exposure time was somewhat longer, i.e., 180 milliseconds ($\frac{9}{50}$ ths of a second). Both studies tested one subject at a time.

The results of Warren's first study reveal that, under the conditions as defined, lower case single-column headlines and all capitals single-column headlines are equally legible. There was an infinitesimal difference of .3 of one per cent but from a statistical point of view and from a practical point of view both kinds of single-column headlines are equally legible at a distance of five and one-half feet. This is the distance one encounters in reading a headline over someone else's shoulder, or across a breakfast table.

In this part of the study, each subject, after completing the test, was asked whether he had noticed any differences in the appearance of the headlines and if so, which kind appeared to be easiest to read. Of the 40 subjects, 25 noted no difference, 8 believed lower case was easier to read and 7 believed upper case was easier to read. From the point of view of the headline reader, therefore, neither kind of headline appeared to have any advantage over the other.

The second study, conducted by Warren and reported in her thesis, was aimed at determining the relative legibility, at various distances, of 100 multi-column or banner headlines when set in lower case and in all capitals.

The headlines selected for study were taken directly from those actually used in Minneapolis papers between 1934 and 1940. They were printed directly on rectangular pieces of cardboard 28" \times 7 $\frac{1}{4}$ " from the largest type immediately available in the University Printing Shop, namely, 60 point size in Memphis bold face.

Since no tachistoscope suitable for exposing such large surfaces was available Miss Warren arranged a stand and exposed each banner headline for approximately one second by removing and replacing a large screen. The group method of testing was employed which permitted determination of the legibility of the two kinds of banner headlines at different fixed distances from approximately 6'2" to 17'1" and at different angles. The student sitting in the center of row one was 6'2" from the exposure stand and those sitting at the end of row one were 7'11". The student sitting in the center of row two was 9'10" from the stand and those at the end of row two were 10'9" away. In row three, the distances

⁷ Warren, A. L. *The perceptibility of lower case and all capitals newspaper headlines*. Master's thesis, University of Minnesota Graduate School, November 1942. On file in University of Minnesota Library.

were 13'4" in the center and 13'8" at the end. In row four the distances were 16'10" and 17'1". Thus, those sitting in row one were glimpsing the headlines at a distance approximating the situation where one looks at a headline across a streetcar, or on a rapidly passed newsstand. On the other hand those sitting in rows two to four were at distances much greater than would ordinarily prevail in daily life. That is, newspaper publishers do not expect that their banner headlines will be read at distances much beyond seven feet. The results, therefore, for row one will be especially significant and should approximate the findings for the first part of Miss Warren's study where each subject was 5½ feet from the tachistoscope.

The results are presented in Table 1. At the first row distance (6'2"

Table 1

The relative perceptibility, at different distances, of five-word 60 point banner headlines set in lower case and in all capitals

Note: Number of lower case headlines = 50; number of upper case headlines = 50. Perfect score would be 250 words correctly apprehended in lower case headlines (5×50) and 250 for upper case headlines. Total number of subjects (N) = 46.

Row	Distance from Row to Exposure Stand		N	Type	Mean Number of Words Perceived	Difference in Per Cent
	Center	End of Row				
1	6'2"	7'11"	12	Lower Case	132.3	- 5.3*
				Upper Case	125.3	
2	9'10"	10'9"	12	Lower Case	117.5	+ 2.6
				Upper Case	120.5	
3	13'4"	13'8"	10	Lower Case	102.4	+ 4.8
				Upper Case	107.3	
4	16'10"	17'1"	12	Lower Case	79.4	+19.8*
				Upper Case	95.2	

* Statistically significant.

to 7'11") the lower case banner headlines seem to be superior to those set in upper case. The loss in perceptibility was 5.3 per cent when the headlines were set in all capitals. This lower case superiority was present for practically all of the students sitting in row one (10 out of 12 made better scores in reading the lower case headlines). The difference is statistically significant at between the 2 per cent and 5 per cent levels of significance.⁸

⁸ This is a technical statistical interpretation based on Fisher's t test of significance and indicates a high degree of probability that the obtained difference would be found to be in the same direction in from 95 to 98 times of 100 repetitions of the experiment.

At first thought one may be inclined to believe that the results of the several studies are inconsistent. The Breland and Breland study revealed a striking superiority of single-column lower case headlines over upper case headlines when perceived at the *normal reading* distance of about 15 inches. The first part of the Warren study showed that both kinds of single-column headlines were equally effective when glimpsed at a distance of $5\frac{1}{2}$ feet. But one should keep in mind that single-column headlines are not intended to be read at any such distance. The second part of Warren's study again discloses the superiority of lower case banner headlines when read at a distance of from six to seven feet which is about the distance at which banner headlines are supposed to be read. Thus, in those situations in which single column or in which multi-column headlines are supposed to be read, it is clear that lower case headlines are distinctly superior to upper case headlines.

The results for rows two and three as given in Table 1 indicate a slight advantage for upper case headlines but the difference is not statistically significant. Furthermore, it is important to note that the average scores for those sitting in rows two and three are lower than the mean score for those sitting in row one. This means that those stationed beyond seven feet from the exposure stand had increasing difficulty in correctly apprehending the headlines even though the exposure time was relatively long (one second).

When we inspect the results for row four in Table 1 it is apparent that the correct apprehension of the five word headlines becomes quite difficult and upper case banner headlines have a distinct advantage. The reading of 60 point banner headlines at approximately 17 feet is thus quite difficult and inaccurate. Apparently the subjects were forced to pick out a word here and there during the one second exposure period. In this sort of situation upper case headlines come into their own since they are larger and the letter outlines are more distinct. This result is quite in line with the earlier study of Tinker.⁹

Our interpretation of the results of the experimental studies now available would lead us to emphasize, as we did originally, the importance of *characteristic word form* which is provided by lower case printing in promoting readability.¹⁰

At least three reasons can be found which will account for the poor legibility of all capitals.

⁹ M. A. Tinker, *op. cit.*

¹⁰ Eye-movement photography in reading all capitals text and in reading lower case shows a 12.4 per cent *increase* in fixation frequency with fewer words perceived for the reading of all capitals. See Tinker, M. A., and Paterson, D. G. Influence of type form on eye movements. *J. exp. Psychol.*, 1939, 25, 528-531.

(1) The printing surface required for upper case headlines is much greater than for the same headlines set in lower case *in the same point size*. A 60 point banner headline is 37.5 per cent longer when set in upper case than when set in lower case. Thus, each visual fixation permits the apprehension of a larger amount of meaningful material in each portion of a headline set in lower case than when the same headline is set in upper case.

(2) The word form is far more characteristic when words are printed in lower case than when they are printed in all capitals. Figure 1 shows that words printed in upper case have a rectangular



FIG. 1. Block outlines of the printed word "stopped" to illustrate that lower case exhibits a characteristic "word form" whereas "word form" is absent when printed in all capitals.

outline, one word differing from another merely in length whereas words printed in lower case or in caps and lower case have an irregular block outline, one word differing from another in a distinctive fashion. This fosters the reading of words as wholes. Figure 2 shows that the

**TO BOLIVIAN DIAN
DADACHIAV OBJECTS**

**Widene in Eritrea
Italiane' Retreat**

FIG. 2. Showing that the upper half of a printed line furnishes more cues to "word form" when printed in lower case.

upper half of a printed line furnishes more cues to "word form" when set in lower case than when set in upper case with the result that its meaning is more readily apprehended.

(3) Reading habits favor lower case since practically all of our reading is devoted to the apprehension of material set in lower case rather than in upper case.

It seems perfectly clear that our original recommendation that upper case printing be abandoned except for an occasional use when one desires

to attract attention through novelty is fully justified. One may now safely conclude that even in billboard printing where reading is to be done at great distances it would be preferable to rely upon caps and lower case than on all capitals. This may seem illogical to some who might be inclined to cite the results reported here as justifying the use of upper case printing for materials to be read at distances beyond seven feet. But it should be pointed out that in all these comparisons we have been observing the effect of using upper and lower case printing *in the same type size*. If one wants banner headlines to be read at a distance greater than seven feet then one need only increase the size of the type from 60 to 72 point or larger and by specifying the use of lower case all of the demonstrated advantages of characteristic word forms as cues to speedy and accurate perception will be retained. One should not forget that larger sizes of lower case than of upper case can be used within a given space. For example, a 60 point banner headline 10" in length in lower case is $13\frac{3}{4}$ " long in upper case. This principle can be extended to car cards, poster signs and billboards which are intended to be read at far greater distances than is true of newspaper headlines. As a matter of fact, we can not think of a single type of reading situation in which upper case printing would be advantageous. A possible exception might be where the perception of an isolated letter to be used as a signal were required as along a railroad or highway. In such rare situations, an upper case letter might be preferred.

Summary

1. The problem of the relative legibility of headlines set in all capital letters and of headlines set in lower case has long been a subject of debate among newspaper men. Scientific studies bearing directly on the problem have been lacking.

2. Several years ago the writers proved that continuous text set in all capitals was read about 12 per cent more slowly than text set in lower case. Generalizing from these results to headlines, the writers recommended the abandonment of all capitals printing. This recommendation was sharply challenged by T. M. Bernstein of the *New York Times* so the writers arranged to have three investigations conducted that would bear directly on the problem at issue.

3. The first study dealt with five word single-column headlines set in 24 point bold face, half in upper case and half in lower case to be read by tachistoscopic exposure at the normal reading distance of 15 inches. The results disclosed an 18.9 per cent difference in favor of the lower case headlines.

4. The second study used the same single-column headlines but exposed them at a distance of five and one-half feet. This time both kinds of headlines were shown to be equally legible.

5. The third study used multi-column or banner headlines set in 60 point bold face which were exposed at varying distances from six feet to seventeen feet. At a distance of six feet the legibility of the lower case banner headline was 5.3 per cent greater than the legibility of those set in upper case. At distances from ten feet to fourteen feet, both kinds of headlines were equally legible. At a distance of seventeen feet, however, the upper case headlines proved to be more legible.

6. The writers conclude that the findings fully justify their earlier recommendation that headline writers should abandon the use of upper case headlines. Caps and lower case printing provides *characteristic word forms* which serve as cues to the rapid and accurate reading of meaningful material. This advantage has now been shown to hold not only for the reading of continuous text but also for the apprehension at a glance of five word single-column and of five word multi-column headlines when exposed at the distances at which they are ordinarily supposed to be read. For reading at unusually great distances such as in billboard display advertising, the writers would insist that lower case is also to be preferred. In this special situation, a larger point size of lower case than of upper case would have to be specified.

7. In conclusion, the writers desire to repeat their earlier recommendation, namely: **THE HEADLINE WRITER AND THE ADVERTISING COPY WRITER SHOULD HENCEFORTH ABANDON ALL CAPITALS and should rely on lower case.** When this rule is generally adopted, an occasional use of all capitals might be justified as a device for attracting attention.

Received March 2, 1945.

Explorations in Personality by the Sentence Completion Method

Amanda R. Rohde

National Hospital for Speech Disorders, New York City

Clinical psychologists, vocational guidance counselors, teachers, and other professional persons who deal with youth problems have continuing need to become intimately acquainted with the needs, inner conflicts, fantasies, sentiments, attitudes, aspirations, and adjustment difficulties of the individuals requiring their services.

Most workers depend largely on the interview and the study of guidance records to discover the intimate traits and tendencies of those whom they counsel. The worker often prefers to supplement the interview with questionnaires, inventories, and personality tests to gain a more complete picture of the individual's traits. Recent years have witnessed a rapid expansion in this field.

All experienced workers realize that the direct question of the interview, the items to be answered or checked in the questionnaires, and inventories have distinct limitations for personality exploration of the deeper sort. Direct questioning tends to make the individual self-conscious; it puts him on the defensive, usually preventing him from disclosing his true self. Freedom of expression is eliminated; the answers are controlled by the questions, the questions in fact suggest the answers.

Fortunately, the so-called projection methods for personality study avoid the resistance that is often met in direct questioning regarding personal matters. Diagnosis by projection techniques reveal latent needs, sentiments, feelings, and attitudes which the subject would be unwilling or unable to recognize or to express in direct communication.

The experimental work with projection techniques has been reviewed by Frank (1, 2) who originated the term, by Murray (6, 7), Horowitz and Murphy (3), Symonds and Samuels (10) and Symonds and Krugman (11).

Projective techniques include: free word association, interpretation of ink blots (Rorschach), picture interpretation (the Murray Thematic Apperception Test, designed to reveal fantasy life; and Stern's Cloud Pictures), autobiography, story completion, story writing, story telling, play techniques with young children, puppet shows, the psycho-drama, music reverie, and expression through the graphic and plastic arts.¹

¹ Descriptions of tests mentioned may be found in the publication by Gertrude Hildreth, *Bibliography of mental tests and rating scales*. New York City: Psychological Corporation, 1939.

The common element in all of these procedures is the diversion of the subject's attention from his own psychic processes and himself as the respondent to the task in hand. Thus he may be led to divulge deep-seated feelings and tendencies of which even he himself may be unaware, and without sensing that he is revealing personal data.

The free word association technique in which the subject is to respond to each stimulus word, e.g. "love" "hate," with the first word that comes to mind, has been widely used for theoretical and practical purposes in determining attitudes, interests, for detecting falsification, emotional states, and conflicts and complexes.

The Sentence Completion Technique

The sentence completion device in which the subject is asked to read to himself the forepart of a sentence and to write anything he wishes to complete the sentence is essentially a projection technique utilizing free association. In unconstrained response to sentence beginnings, the subject inadvertently reveals his true self, since there is no way in which he can anticipate the significance of his answers for personality study.

In 1928 Payne (8) published a sentence completion test of fifty items that was widely used in vocational counseling. Payne devised the test for use with college students to reveal personal traits through eliciting inhibited responses.

Tendler (12) has developed a twenty-item sentence completion test in which each item except the first begins with "I," for example: I feel happy when_____.

Lorge and Thorndike (4, 5) experimented with a 240-item sentence completion test in which the subject was restricted to single-word responses. They reported negative results in using the device for personality study, since they found little relation between test responses and the actual behavior of the subjects tested. These results are not surprising since the subjects were restricted to single-word replies; and judging from the samples of items given in the report, the items were none too carefully worded for the intended purpose. The device appears to be better suited, as is the classical free word association method, to lie detection and to the discovery of specific objects about which mental complexes center. In a single-word response the subject's expression is too highly controlled for him to reveal much about himself no matter how many items are used.

Sanford *et al.* included a "Completion of Sentences Test" among other projection techniques in an extensive study of children (9). The test presented the subjects with a series of 30 incomplete sentences—clauses or phrases—so constructed as to permit a wide variety of gram-

matical conclusions. By selecting systematically the content of the stimuli it was possible to limit the range of responses provided the subjects followed the rules of grammar in their answers. The intention was to explore certain response areas. The following classes of stimuli were included. (1) A series of press. The given phrase or clause described an environmental situation or event affecting some actor and was so arranged that a grammatical completion gave an account of the actor's response, an action pattern which could usually be classified according to the scheme of needs. (2) The first part of the sentence described an action pattern of some need and was so arranged that to complete the sentence grammatically was to give the situation or event (press) which aroused the need. (3) The phrase or clause given was ambiguous in that the action was incomplete or lacking in motivation and so constructed that a grammatical completion supplied the missing motive (need).

The test was administered to children in grades 3 to 9 inclusive. The results were combined with those of a "Completions of Pictures Test" and an "Interpretations Test" since it was concluded that not any one of these by itself could add much to the knowledge of the general project. Thus, regarding quantitative results, each of the tests was considered to remain in an exploratory stage of development. In so far as qualitative results were analyzed, the answers revealed consistent patterns of response as well as responses of no significance. However, the difference of response—even of the same stimulus material—indicated fairly clear differences of personality among the children tested.

Revision and Extension of the Payne Sentence Completion Blank

From experience with the original Payne Sentence Completions Blank, the writer became convinced that the method merited wider use as a projective technique for studying the personal characteristics of young people. In 1939, having obtained permission from the author's heirs for this purpose, a revision and extension of the Payne completions blank was worked out. The aim was to develop an instrument which would have the advantages of other projective techniques, but simplicity of procedure and interpretation to give practical value for psychological diagnosis in schools and other institutions where the study of large numbers of individuals is required.

Considerable preliminary experimentation was required to select the proper range of items, to phrase the items, and to arrange the order of items within the test so that the minimum number of items would be included that would yield maximum indications for diagnosis.

The criteria used in selecting and constructing items were as follows: (1) The range of different stimuli must be broad enough to elicit informa-

tion concerning all phases of personality. (2) The responses must be controlled as little as possible by the stimulus phrases so that the subject may have freedom of expression. (3) The total time required for the test must not exceed a period convenient for schedules of schools and institutions.

Among the items finally retained were some that were freer of control than others. However, they were on the whole somewhat less controlled than Payne's original items. Of these, thirty-five were retained in original form, four in modified form; the remainder were dropped and twenty-five new ones added. The final test consists of sixty-four items.² The following are illustrative sentence beginnings used in the questionnaire:

My school work.....
 I want to know.....
 There are times.....
 My greatest longing.....
 Girls.....
 Earning my living.....
 At night.....
 I cannot understand what makes me.....
 My father.....
 When I.....
 Death.....
 My eyes.....
 Love.....
 At home.....
 I become embarrassed.....

Following the last item there is a blank space with the instructions: Write below anything that seems important to you.

The Rohde-Hildreth Sentence Completion Blank is arranged as a four-page folder. On the cover page, space is reserved for identification data and for a summary of the test findings.

A code system may be used if the examiner prefers to instruct examinees not to write their names on the blanks. When the test is administered to groups, more freedom of expression is usually obtained if this procedure is adopted. It is recommended that the questionnaire be used during a regular class period, preferably an English class. The examiner instructs the students to complete the sentences that are partially begun and informs them that any response they care to make to any item will be entirely acceptable. If the blank be used for adults in other situations, it usually suffices merely to say that it is a sentence completion exercise.

There is no time limit. The subject proceeds at his own pace. Some people finish in half an hour; others require an hour and a half. The median time is 45 to 50 minutes.

² Sentence Completions by Amanda R. Rohde and Gertrude Hildreth, copyright 1940.

The questionnaire is intended for use with individuals approximately twelve years of age and above, including adults. Its use as a group device does not invalidate the results for individual diagnosis.

A standard scoring scheme has been worked out by means of which the results may be expressed in quantitative form.³

While the subject's needs (motives), inner states (conflicts, dejection, optimism, etc.), traits, and the press (environmental forces) revealed are the chief bases of quantitative results, many other important aspects of personality are manifested. One gets a fairly good idea of his tastes and sentiments, something of his ideology or philosophy, and the general representation of the ego structure, besides indications of intellectual status and emotional maturity.

Incidentally, the subject's language fluency, diction, grammar, ability to spell and punctuate, as well as the legibility and character of handwriting may be taken into consideration in evaluating the responses of the individual.

Compared with other projection devices, the sentence completion blank is time saving in two ways: First, it may be administered effectively to groups of persons at one time. Secondly, even if it be used individually, after the initial directions are given, the instructions make it self-administering.

Whether or not a subject responds "truthfully" is immaterial from the standpoint of exploring the individual's traits and tendencies. Repressed feelings, attitudes, and ideas emerge regardless of whether the subject be "telling the truth" or not. One subject commented: "I've answered this as though I were another person." This fact did not vitiate the results as one might be inclined to assume; on the contrary, this significant comment only added to the validity of the personality picture furnished by the test responses.

The fact that an individual student may suspect the purport of the test does not invalidate the technique. Attempts at deception, false or cryptic answers of suspicious persons, are not less revealing than those of others who give full, free responses; because each person always writes what to him seems a good response, thus projecting his personality regardless of his intentions.

The sentence completion device has proved to be a valuable adjunct to the interview and in the advisement connected with vocational guidance and orientation courses. It is especially recommended for the study of persons who present adjustment problems.

³ A manual with directions for administration and scoring has been prepared by the writer. Copies of the test and manual may be obtained from the writer, 411 West 115th St., New York City, 25, N. Y.

Although the device has been used successfully by teachers and guidance workers in understanding individual adjustment problems, the most penetrating analyses of results are possible only in the hands of clinicians with extensive psychological training and experience.

Validation of the Rohde-Hildreth Sentence Completion Technique

Since 1939 the revised sentence completion test has been used experimentally by the authors and selected workers, chiefly with students of high school age. It has also been used to study the personal adjustments of patients in the National Hospital for Speech Disorders of New York City, of college students, of public health nurses, and it has been tried out experimentally by psychologists in one of the U. S. Navy Hospitals.

In appraising the Rohde-Hildreth device, the questions to be answered are: Does the test serve the function it was intended to serve? Is it useful or helpful in any way in studying or appraising the individual? Is it a reliable instrument?

The writer has done the following described experimental work for test validation. The subjects chosen for the validation study were ninth grade high school students in public schools, including altogether 670 students from several high school populations. This grade was chosen because at this age level students are in a critical stage of personal adjustment, and understanding of the adolescent's personality, problems, conflicts, and aspirations is a highly important function of teachers and counselors. The average age of these pupils was fifteen years.

Murray's conceptual scheme of personality (7) provided the essential concepts of dynamic behavior units and variables for interpretation and evaluation of experimental data obtained with the completion test. All test blanks were analyzed in terms of the Murray categories.

The questionnaire responses were referred to a list of 39 personality categories based on the Murray schema, including such items as abasement, aggression, creativity, deference, dominance, flight from danger or harmavoidance, passivity, rejection, seclusion, sex, conflict, emotionality, optimism, and Superego (conscience).

Using this classification and description of variables as a guide, it was possible to translate the sentence completion data into quantitative results. This was accomplished, considering the test as a whole, first by analyzing the contents of the sentence completions for each individual into needs, inner states, and press; and secondly by determining their frequencies and intensities. Frequency and intensity in a given subject's responses were estimated as follows: Frequency was obtained merely by counting the number of occurrences of each variable. In

estimating intensity, a one to three scale was used: 1. low, 2. average, 3. high, the criteria for judging being the vividness and potency with which a variable was expressed.

For purposes of scoring, it was deemed best to score only gross categories and the main subdivisions: the major forms of needs, inner states, and press.

From the following brief excerpts of the papers written by two ninth grade boys, designated "A" and "B," one may gain some idea of the method of evaluating and of scoring the responses. One may also obtain a fairly good notion of the personality of the boys, although less than one-third of their answers are quoted. The varying needs and press indicated by responses to the same sentence beginnings are shown since similar numbers were extracted from each paper.

"A's" paper was written by an orphan boy who had lost both parents. He lived in a children's home in New York City and was well thought of by his teachers and school mates. He studied hard, was much interested in art, but disliked algebra. His marks in art, English and social studies were above average, while he managed to get barely a passing grade in algebra. After graduation from the junior high school, he was admitted to the School for Music and Art, a senior high school, where he pursued his work with keen enjoyment. It is significant that his great need for adventure, excitement and change impelled him to enlist and join the armed forces three months before graduation when he learned that diplomas would be granted to the boys who had fulfilled the requirements up to that time. Thus, further follow-up was prevented.

Subject "A"

Age: 15 yrs. 7 mos.

I.Q. 114

1. I want to know if all have the same feeling for art as I do.
2. The future seems very bright and cheerful.
3. My school work has been very interesting to me.
4. Earning my living is a thrill.
5. My greatest longing is to paint.
6. Secretly I steal food from the pantry.
7. If I fail in algebra, I would practically "die."
8. There are times when I feel like running away and start a new life.
9. Work is pleasant and hard.
10. Friends are a help and encouragement.
11. I become embarrassed when I can't dance.
12. Girls fascinate me.
13. Love is grand!
14. Other people are quite interesting.
15. The laws we have are sometimes unjust.
16. I cannot understand what makes me so nervous and stammer.
17. My stomach is fine and holds a lot.
18. At night I study.
19. My mother is dead!
20. Death is sometimes inviting.

An analysis ^a of the above responses gives one the impression of ■ worried, mentally stimulated, but yet relatively normal and adjusted boy who has held his own despite some severe handicaps. He appears to be slightly below average in scholarship. His preference for art as compared with algebra marks him as one with a verbal or pictorial mind rather than mathematical intelligence, and one who would not get universally good grades. The fact that his attitude as expressed here is amiable and optimistic, despite his obstacles, suggests a fairly well-structured ego, built up by pride and reinforced by a satisfactory conscience.

The outstanding press (environmental forces) indicated are:

- p Affiliation—Sentence Nos. 10, 12, 14.
- p Aggression (Belittlement or ridicule, possibly on account of his stuttering, inability to dance, and failures in algebra, plus some inferiority in regard to his speech)—Sen. Nos. 7, 11 and 16.
- p Dominance—Sen. No. 15.
- p Imposed task—Sen. Nos. 7, 9, 18.
- p Loss of support (Death of mother and perhaps lack of nurturance from sub-parents)—Sen. No. 19.

The most important needs (n) and traits expressed are:

- n Abasement (Masochism, thoughts of suicide, and leaving the field of failure)—Sen. Nos. 7, 8 and 20.
- n Achievement (Interest in schoolwork, earning a living, and painting) fused with a good deal of n Counteraction (Desire to overcome weaknesses and obstacles)—Sen. Nos. 3, 4, 5, 9, 18.
- n Acquisition (Desire to earn money, and stealing food out of lack)—Sen. Nos. 4, 6.
- n Autonomy and freedom (Motility, change, excitement) fused with n Inviolacy (Intolerance of failure—i.e., leaving field of failure), also with n Cognizance, n Abasement and n Harmavoidance (low)—Sen. Nos. 8 and 15.
- n Affiliation fused with some n Succorance—Sen. Nos. 10, 12, 14.
- n Cognizance (Curiosity and desire for knowledge, interest in schoolwork, travel, and new scenes) fused with n Autonomy and freedom—Sen. Nos. 1, 3, 8 and 14.
- n Harmavoidance (Low, repressed by pride)—Sen. Nos. 7 and 8.
- n Inviolacy (Pride and intolerance of belittlement and failure) fused with n Infavoidance (Desire to escape shame and humiliation)—Sen. Nos. 7, 8, 11 and 16.
- n Nutrition—Sen. Nos. 6 and 17.
- n Sentience (Interest in beauty, color and other sensuous impressions)—Sen. Nos. 1 and 5.
- n Sex—Sen. Nos. 12 and 13.

The outstanding trait and inner states distinguished are: Emotionality—Sen. Nos. 1, 2, 4, 8, 16. Optimism and cheerfulness—Sen. Nos. 2, 4, 9, 10 and 17.

^aThis analysis is ■ partial quotation from the blind analysis made for the writer by H. A. Murray who also gave other valuable suggestions in connection with this study.

The Ego Ideal is to become an artist—Sen. Nos. 1 and 5.

The positive cathexes (power of objects to arouse responses of a certain kind in subjects) expressed are: art, food, friends, girls, painting and sometimes death. The negative cathexis is algebra.

The personality profile given in part above was corroborated, as were those of other subjects studied, by interviews with the subject himself, with teachers, administrators, schoolmates, and others who had been associated with the boy.

In contrast with the frank, cheerful, optimistic person just discussed, we find in the following responses of subject "B" an attitude of dark pessimism, envy, distrust and antagonism, a tendency to make idealistic statements that have little support, and to do much wishful thinking rather than striving to achieve his ambitions as indicated by an analysis of the needs, inner states and press.

Subject "B"

Age: 14 years 3 months

I.Q. 118

1. I want to know *why mankind is destroying itself.*
2. The future *is an infinite darkness.*
3. My school work *has been improving of late.*
4. Earning my living *by science is my greatest ambition.*
5. My greatest longing *is to master science.*
6. Secretly *I hope I will improve physically.*
7. If I *can become an optometrist, my ambition will be realized.*
8. There are times *when I wish I already had a full knowledge of science.*
9. Work *is unfair under all present regimes.*
10. Friends *are a beauty of nature.*
11. I become embarrassed *(not answered).*
12. Girls *are better companions than boys.*
13. Love *(not answered).*
14. Other people *are interesting specimens of personality.*
15. The laws we have *are the best in the world but still not perfect.*
16. I cannot understand what makes me *so poor in spelling even if I study two hours straight.*
17. My stomach *don't agree with sardines.*
18. At night *I journey into the vast beyond in wild adventure.*
19. My mother *loves my brother more than me.*
20. Death *is a great adventure.*

The outstanding needs (n) and traits are:

- n Achievement—Sentence Nos. 3 and 16—fused with n Counteraction (Desire to overcome obstacles)—is not sufficient to support his Ego Ideal (high aspirations)—Sen. Nos. 4, 5, 7 and 8.
- n Affiliation is fused with n Fantasy—Sen. No. 10—and becomes empty sentimentality when considered with Sen. No. 12 which expresses a fusion of n Affiliation, n Aggression (Jealousy of possible rivals among boys) and n Sex. (Cf. also n Aggression.)
- n Aggression (Envy of brother—Sen. No. 19—and antagonism towards social-economic conditions—Sen. No. 9—plus the Narcissistic attitude towards other people—Sen. No. 14) further points to lack of sincerity in friendship.

- n Cognition (Desire for knowledge) is high—Sen. Nos. 1, 4, 5, 7 and 8—but is stagnated by n Fantasy (Desire to escape from reality). (Cf. n Fantasy and sentence references below.)
- n Counteraction (Desire to overcome weakness) is concerned with his physical condition rather than moral weaknesses—Sen. No. 6.
- n Fantasy and irreality (Desire to dream or to do autistic thinking)—Sentence Nos. 8, 18 and 20. The desire for perfectionism—Sen. No. 15—is a further manifestation of n Fantasy which here is fused with n Deference (Praise of laws).
- n Seclusion and secrecy (Failure to answer Sen. Nos. 11 and 13) also suggest evasion and lack of sincerity.
Pessimism and dejection probably are a projection of bitterness resulting from antagonism towards his mother, jealousy of his brother and other boys, and self-centered attitude toward people and the world in general—Sen. Nos. 1 and 2.

The important press (p) are:

- p Lack of support (Mother's preference for brother)—Sen. No. 19.
- p Foreboding anxiety—Sen. Nos. 1 and 2.
- p Imposed task—Sen. Nos. 3 and 16.
- p Physical affliction—Sen. No. 6.

The positive cathexes are: science, friends, girls, and death.

The negative cathexes are: work, present regimes, boys, sardines, mother, brother and probably spelling and schoolwork. The latter it may be noted outweigh the former.

Interviews with teachers and supervisors disclosed that the boy was considered brilliant but lazy, therefore made poor grades in most of his subjects with the exception of one or two things which he especially liked. In a personal interview he criticized all of his teachers, his mother and brother, and all social and economic conditions. There was no evidence of ill-health. On the contrary, he looked strong and healthy. Some tactful questioning brought forth the information that the great desire to build up physically stemmed from self-admiration and the wish to be attractive to the opposite sex. He was reported to be indulged by his parents. In consequence of lack of understanding and pampering, the boy had developed thinking habits which prevented satisfactory progress in school.

To translate the judgments of teachers, counselors, and the experimenter into quantitative terms, the rating for every variable for each student was determined on a scale of one to ten. The score for each variable was then found by averaging the independent ratings of the judges.

To determine the validity of these analyses and the interpretations, the scores obtained from random samples of test paper responses of fifty girls and fifty boys were correlated with the combined ratings obtained from teacher's judgments of the subjects, the experimenter's interview

Table 1

Showing Correlation Coefficients between Ratings of Student's Responses and Ratings of Combined Judgments of Teachers and Other Sources Previously Mentioned (Columns 1 and 2), and between Ratings of Initial and Final Tests (Columns 3 and 4—Corrected for Attenuation)

Needs	Girls	Boys	Girls	Boys
Abasement	.89	.89	.97	.70
Achievement	.49	.59	.85	.71
Acquisition	.87	.75	.55	.77
Affiliation	.65	.77	.91	.85
Aggression	.85	.85	.56	.62
Autonomy	.90	.90	.72	.70
Blame avoidance	.81	.83	.74	.63
Change	.96	.94	.87	.73
Cognition	.92	.92	.94	.95
Counteraction	.53	.54	.75	.47
Deference	.85	.76	.82	.82
Defendance	.93	.72	.85	.82
Dominance	.30	.53	.65	.82
Exhibition	.78	.74	.90	.93
Exposition	.60	.87	.91	.98
Fantasy	.94	.89	.79	.84
Infavoidance	.95	.86	.81	.64
Nurturance	.69	.88	.88	.84
Organization	.53	.85	.84	.97
Playmirth	.63	.86	.85	.74
Recognition	.86	.75	.84	.80
Seclusion	.85	.93	.97	.94
Sentience	.69	.91	.90	.67
Sex	.69	.80	.84	.44
Succorance	.88	.83	.50	.84
Inner States				
Conflict, Perplexity	.82	.91	.76	.39
Dejection, Pessimism	.93	.88	.85	.56
Exultation, Optimism	.93	.93	.92	.63
Emotionality (Trait)	.89	.86	.72	.81
Inner Integrates				
Ego Ideal Achievement	.90	.90	.58	.75
Ego Ideal Pride	.91	.82	.89	.78
Superego	.88	.96	.74	.80
Narcism	.72	.80	.83	.84
Averages*	.79	.82	.82	.76

* The difficulty in securing adequate validation of such variables as Nutriance, Passivity and Rejection by teacher ratings necessitated the omission of these variables in the table. These variables were included, however, in the computations to determine consistency of response, but are not indicated in the table shown above which is an abridgment of the four original tables. Averages given in columns 3 and 4 were taken from the original table.

with them and their parents, the opinions of school administrators and guidance counselors concerning the students, and from their school records.

These scores and the test paper scores were converted to equivalent scores for computation of correlation coefficients.

The linear correlations (Pearson product-moment) between ratings of students' responses in the sentence completions items and the ratings of the combined judgments of teachers, the experimenter's interview ratings and other sources previously mentioned, all categories combined, were as follows: 0.79 for the girls, and 0.82 for the boys. For the different categories the correlations varied from 0.95 to 0.295. The lowest correlation was found for the need dominance. The result suggests that the test may not provide adequate stimuli for the expression of this need. Table 1 shows r 's for variables.

To obtain an estimate of the reliability of scoring the sentence completion responses, the scorings of from one to four raters in addition to the original experimenter were obtained. All these judges were thoroughly familiar with Murray's conceptual scheme of variables. There was a 95.5 per cent agreement in the items scored by the experimenter and an independent rater on 36 papers, while there was a 78 per cent agreement when five judges rated the same items on twelve papers.

Consistency of responses for 21 girls and 23 boys was retested after eight months. The averages of obtained r 's between initial and final test scores, corrected for attenuation by the Spearman-Brown formula, were 0.82 for the girls, and 0.76 for the boys.

The greatest changes after eight months were found in the individuals who were not well adjusted emotionally and socially. Apparently the unadjusted person who has unresolved problems strives unceasingly, first in one way then another, to find solutions. The needs of the adjusted person appear to be more stable.

It is impossible to give here a full account of the experimental results, to indicate the typical problems, fears, wishes, and difficulties that characterize the ordinary ninth grade high school students.

The study revealed that teachers seemed to be capable in observing objective needs arising from environmental conditions. Deep-seated emotional conflicts were less readily understood. In some cases maladjusted students were considered queer, but the underlying causes were not suspected by parents or teachers.

Thirty per cent of the girls and 32 per cent of the boys expressed concern about passing in their studies. A large per cent had worries connected with school. The greatest longing expressed by these students was to attain their Ego Ideal. Doing something interesting or thrilling was the greatest longing of 20 per cent of both boys and girls.

References

1. Frank, L. K. Projective methods for the study of personality. *J. Psychol.*, 1939, 8, 389-413.
2. Frank, L. K. Projective methods for the study of personality. *Trans. N. Y. Acad. Sci.*, 1939, 1, 129-132.
3. Horowitz, Ruth, and Murphy, Lois B. Projective techniques in the psychological study of children. *J. exp. Educ.*, 1938, 7, 133-140.
4. Lorge, I., and Thorndike, E. L. The value of the response in a completions test as indications of personal traits. *J. appl. Psychol.*, 1941, 25, 191-199.
5. Lorge, I., and Thorndike, E. L. The value of responses in a free associations test as indicators of personal traits. *J. appl. Psychol.*, 1941, 25, 191-199; 200-201.
6. Murray, H. A. Techniques for a systematic investigation of fantasy. *J. Psychol.*, 1936, 3, 115-143.
7. Murray, H. A., et al. *Explorations in personality*. New York: Oxford University Press, 1938.
8. Payne, A. F. *Sentence completions*. New York Guidance Clinic, 1928.
9. Sanford, R. N., et al. Physique, personality and scholarship. Washington, D. C. *Soc. Res. Child Devel., National Research Council*, 1943.
10. Symonds, P. M., and Samuel, E. A. Projective methods in the study of personality. *Rev. Educ. Res.*, 1941, 11, 80-93.
11. Symonds, P. M., and Krugman, M. Projective methods in the study of personality. *Rev. Educ. Res.*, 1944, 14, 81-98.
12. Tandler, A. D. A preliminary report on a test for emotional insight. *J. appl. Psychol.*, 1930, 14, 123-136.

Speech Intelligibility Under Various Degrees of Anoxia *

G. M. Smith and C. P. Seitz †

College of the City of New York

This study is an extension and a refinement of an earlier study by the same investigators (3). In the earlier study it was demonstrated that an altitude of 18,500 ft., simulated in a nitrogen dilution chamber, produced a statistically reliable decrement in speech intelligibility under certain conditions of initial difficulty. It was pointed out that the magnitude of the decrement, though not large, was indicative of a change that, under trying conditions, could become hazardous in commercial and military aviation. Though it is common practice to use oxygen above 10,000 or 12,000 ft., a pilot flying at 40,000 to 45,000 ft. breathing pure oxygen has, because of the reduced oxygen tension in the lungs, a physiology which roughly corresponds to that of a man flying at 16,000 to 18,000 ft. without oxygen. The demonstration of a decrease in speech intelligibility in this general altitude range is, therefore, not without some practical significance.

It was felt that the effect of altitude on speech intelligibility observed in the previous study would have been greater had some more adequate means been found to equate the difficulty of the test materials used in the altitude runs with those used in the control runs. It also seemed desirable to investigate the possibility of a decrement in speech intelligibility at altitudes somewhat lower than 18,500 ft., the altitude selected in the first investigation. The purpose of the present study was to make such an investigation with improved techniques.

Method and Materials

The method of observing the effect of oxygen deprivation in the present study involved the subjects' ability to hear and to indicate on a check list words in common speech. The use of a check list by the subjects was an innovation, for in the earlier study the subjects reported to the experimenter who checked their responses for them. The new pro-

* The authors are indebted to the Linde Air Products Co. for a liberal grant of oxygen and nitrogen, and to Messrs. Mortimer Feinberg and Max Rosenbaum for valued clerical and statistical assistance.

† On leave, Lieutenant U.S.N.R. The opinions or assertions contained herein are the private ones of the writers and are not to be construed as official or reflecting the views of the Navy Department or the Naval Service at large.

cedure eliminated the possibility of errors being introduced by the experimenter. The possibility that errors on the subjects' check list might be errors in visual perception and in checking rather than errors in auditory perception was minimized by the use of a procedure described below.

Uniformity in the difficulty of the test materials used during the control and the three experimental sessions was assured by the use of eight lists of recorded stimulus words. Test batteries used for any one of the four conditions were made up of the same lists of recorded test words, but they were presented in varying order. The test series were made up in part from standard word lists developed by the Bell Telephone Laboratories (2), covering the more frequent sounds that occur in common speech. Other test series were constructed by the investigators on the principles employed by the Bell Laboratories. Intelligibility for vowel sounds was tested by the use of lists of monosyllables all having the same initial and final consonants in any one list; e.g., *suit*, *sit*, *sat*, *set*, etc. Consonant intelligibility was tested by similar lists, each involving a constant vowel sound but a variation in the initial or final consonant; e.g., *nor*, *bore*, *yore*, *more*, etc. In each list there were 11 vowel items and 24 consonant items.

The recordings of the test materials were made on high fidelity equipment at the National Broadcasting Co. Studios in New York City.¹ To minimize the effect of wear these recordings were later put in semi-permanent form by the RCA Manufacturing Co. of Camden, N. J.² The recordings were played back through a Fairchild pick-up³ coupled with a Presto amplifier⁴ (especially adapted so as to give a relatively flat response curve) and Western Electric earphones.⁵

The experiments were carried out in the nitrogen dilution chamber of the College of the City of New York, which has a capacity of approximately 450 cu. ft. and walls approximately 8 in. thick, which provided adequate sound-proofing. An air conditioning unit maintained constant and comfortable temperature and humidity conditions, 74° F. and 60% relative humidity. The three simulated altitudes were 13,600 ft., 16,900 ft., and 20,100 ft. (corresponding to oxygen percentages of 12.5, 10.3, and 8.85, respectively). Samples of the chamber air were taken after altitude had been attained in each run and were analyzed on the Haldane-Henderson-Bailey gas analysis apparatus. The samples did not deviate from

¹ We are indebted to the National Broadcasting Co. and to Mr. R. A. Lynn of the engineering department for their cooperation.

² We are indebted to the RCA Manufacturing Co. and to Mr. W. L. Tesch of the record engineering department for their courtesy.

³ Turntable unit model 199 and pick-up model 209.

⁴ Model 87B.

⁵ Type 588A.

the desired altitudes by more than 600 ft., with one exception.⁶ The CO₂ content of the chamber air averaged 0.52%.

Subjects

Twelve male college students free from systemic defects as determined by medical examination served as subjects. They were supervised during each experimental run by a physician⁷ in addition to the experimenter. The ages ranged from 18 to 21 years, with the median at 19 years.

Procedure

The subjects worked in four groups of three, each group being tested under four separate conditions: sea level and three altitudes, 13,600 ft., 16,900 ft., and 20,100 ft. On the sea level run all the usual routine for altitude runs, including the use of an oxygen mask by the experimenter, was followed; so the subjects were not aware that it was a control run. The usual precautions to allay fear of the chamber were taken. During the testing subjects were comfortably seated and were equipped with standard Western Electric earphones such as are used by the American Airlines on transport planes.⁸ Record booklets made up of twelve separate tests, each containing check rows for eleven vowels and twenty-four consonants, were employed.

Table 1
Test Order

Test Number	Sound Level	Condition
Instructions	High (30 db.)*	Sea Level
1	High	Sea Level
2	High	Altitude**
3-10 (main series)	Low (24 db.)***	Altitude**
11	High	Altitude**
12	High	Sea Level

* Vowel articulation was approximately 98% and consonant articulation was approximately 94% at sea level.

** On control runs this was a pretense only.

*** Vowel articulation was approximately 92% and consonant articulation was approximately 51% at sea level.

⁶ During part of one of the four 16,900 ft. runs the altitude dropped to 15,800 ft.

⁷ We are indebted to Dr. William Recht for medical supervision.

⁸ Type 588A. The response curves of one set of phones were plotted at the Stevens Institute of Technology through the courtesy of President H. N. Davis and Dr. H. Burris-Meyer. Though the curves are peaked rather sharply at 1,000 cps., they are relatively flat (± 10 db.) between 2,000 and 8,000 cps. The transmission system as a whole was sufficiently free from distortion to make possible a ready identification of the speakers' voices. We are indebted to the American Airlines and to Messrs. D. W. Rentzel and H. A. Wolfe for the use of the phone sets.

The order and condition of test administration is indicated in Table 1. The first test, following the recorded instructions, was given at sea level at a sound intensity level so high that it offered almost no difficulty in intelligibility (see first footnote under Table 1). The second test was administered at altitude ⁹ at the same high sound level, after a 15 minute period of adjustment. Then the main series of tests, tests 3-10, followed at a much lower sound level, intentionally set so that the consonant articulation (the percentage of correct responses) was in the neighborhood of 50% (see third footnote under Table 1). Following the main test battery, the sound level was again increased to the initial high value and test 11 was taken while still at altitude. Finally, following return to sea level, test 12 was administered at the same high sound level. The purpose of these initial and final tests at the high sound level was to provide a check on the possibility of errors due to the subjects' inability to manipulate pencils accurately or to perceive the response on the check list in the five seconds allowed for this purpose.

The experimenter remained in the chamber throughout the run to supervise the procedure, to check on the condition of the subjects, and to allay any apprehension. The average time of ascent was 14 minutes, the time for the administration of the tests was approximately 45 minutes; thus, the total time in the chamber, including the 15 minute period of adjustment, was approximately 75 minutes.

The test booklets were in four forms, one for each of the four sessions. Identity of difficulty with a minimal practice effect was achieved by varying the order in which the recordings were played back. The order in which the four different conditions in the chamber occurred was different for each of the four groups. This was arranged in such a way that any advantage attributable to order per se was reduced to a minimum.

Results

The principal data are summarized in Table 2 which gives the articulation values for vowels, consonants and standard syllables for each of the twelve subjects, at sea level and at the three altitudes, for the main series of tests (tests 3 to 10 combined), which were administered at the low sound level. Standard syllable articulation values were calculated by the Fletcher and Steinberg formula $S = 1 - (1 - VC^2)^{0.9}$ derived empirically from extensive observations in the Bell Laboratories. The means and the probable errors of these means are also presented in this table. As is well known, the articulation values for vowels are very much better than for consonants under all conditions. The mean values for vowels, consonants and standard syllables all drop systematically as the altitude

⁹ The ascent was of course a pretense on the control runs.

Table 2

Articulation Values for Vowels, Consonants, and Standard Syllables
Tests 3 to 10 Combined* (Low Sound Level)

Subject	Sea Level			13,600 Feet		
	Vowel	Conso- nant	Syllable**	Vowel	Conso- nant	Syllable**
1	93.0%	49.5%	21.0%	84.0%	47.0%	17.0%
2	100.0	63.5	37.0	96.5	49.0	21.0
3	70.5	28.5	5.0	35.0	21.5	1.5
4	97.5	63.5	36.0	85.0	42.0	13.5
5	94.0	46.0	18.0	90.0	44.5	16.0
6	99.0	63.0	36.0	96.5	70.5	44.5
7	79.5	34.0	8.5	86.5	49.0	19.0
8	96.5	68.5	42.0	67.0	52.0	16.5
9	94.0	42.0	15.0	100.0	69.5	44.5
10	97.5	64.0	37.0	100.0	57.5	30.5
11	82.0	31.5	7.5	88.5	40.0	12.5
12	95.5	60.5	32.0	100.0	49.5	22.5
Mean	91.6	51.2	24.6	85.7	49.3	21.6
P.E. _M	1.7	3.1	3.0	2.9	1.9	2.3

Subject	16,900 Feet			20,100 Feet		
	Vowel	Conso- nant	Syllable**	Vowel	Conso- nant	Syllable**
1	71.5%	37.5%	9.0%	35.0%	14.5%	0.5%
2	97.5	49.5	22.0	36.5	26.5	2.5
3	65.0	29.5	5.0	57.0	25.0	3.0
4	63.5	25.0	3.5	54.5	29.5	4.5
5	41.0	17.0	1.0	48.0	16.5	1.0
6	68.0	25.0	3.5	54.5	19.5	2.0
7	63.5	15.5	1.5	17.0	2.0	0.0
8	51.0	21.0	2.0	7.0	0.0	0.0
9	48.0	8.5	0.5	13.5	0.5	0.0
10	99.0	60.5	33.5	94.0	52.5	23.5
11	52.0	26.0	3.0	41.0	12.0	0.5
12	90.0	37.5	11.5	61.5	35.0	7.0
Mean	67.5	29.4	8.0	43.3	19.5	3.7
P.E. _M	3.6	2.7	1.8	4.5	2.9	1.0

* Total number of test items equals 278 for each subject (190 consonants, 88 vowels).

** $S = 1 - (1 - VC^2)^{0.9}$.

increases. In the case of syllable articulation the mean value at sea level is 24.6%; at 13,600 ft., 21.6%, a negligible drop. But at 16,900 ft. a sharp decrement has occurred, bringing the mean to 8.0%, and at 20,100 ft. the mean has dropped to 3.7%. The values for vowel and consonant articulation show a similar trend.

These relations are graphically portrayed in Figure 1, which makes it clear that somewhere between 13,600 ft. and 16,900 ft. there is a critical value for the particular order of difficulty of intelligibility used in this experiment. It should be emphasized that the character of this curve is very definitely a function of the initial difficulty at sea level. This statement is supported, not only by the earlier experiment, but by the supplementary tests in the present experiment. In the earlier study, where the initial difficulty was less, the mean syllable articulation at sea level (55.5%) dropped only 8% at 18,500 ft., as compared with decrements of 16.6% at 16,900 ft. and 20.9% at 20,100 ft. in this study (tests 3-10 combined), where the mean syllable articulation at sea level was 24.6%.¹⁰ In the supplementary tests of this study (tests 1, 2, 11, and 12) administered at the high sound level, the decrement at altitude is also much less. This is indicated in the bottom row of Table 3, which compares the mean

Table 3
Comparison of Mean Syllable Articulation Values

Means Compared	Tests Involved	Sound Level	Difference between Means	<i>t</i> Values*	<i>P</i> (Probability that difference is due to chance)**
S.L.-13,600 ft.	3 to 10	Low	24.6-21.6 = 3.0%	0.76	.46
S.L.-16,900 ft.	3 to 10	Low	24.6- 8.0 = 16.6%	4.5	.001
S.L.-20,100 ft.	3 to 10	Low	24.6- 3.7 = 20.9%	5.7	.0002
13,600-16,900 ft.	3 to 10	Low	21.6- 8.0 = 13.6%	3.1	.010
16,900-20,100 ft.	3 to 10	Low	8.0- 3.7 = 4.3%	2.6	.025
20,100 ft. loud- 20,100 ft. soft	2+11*** vs. 3 to 10	High- Low	57.9- 3.7 = 54.2%	8.6	<.00001
S.L. loud- 20,100 ft. loud	1+12*** vs. 2+11***	Both High	65.3-57.9 = 7.4%	1.1	.29

* These values were calculated by R. A. Fisher's technique (1) for testing the reliability of a difference between the means of small correlated samples.

** *P* was calculated by Fisher's method from Student's tables of *t*. The difference may be regarded as reliable when *P* is .01 or less.

*** These two tests combined involved only 22 vowel and 48 consonant test items, as compared with the combination of tests 3-10, which involved 88 vowel and 190 consonant items.

¹⁰ The more marked decrements in this study are in part due to the more sensitive procedures employed.

syllable articulation for tests 1 and 12 taken at sea level with the mean for tests 2 and 11 taken at 20,100 ft. Here the decrement of only 7.4% is clearly due to the fact that the initial difficulty at sea level was relatively slight, as indicated by a mean syllable articulation value of 65.3% in contrast with the mean value of 24.6% in the main series.

Table 3 verifies the impression given in Figure 1. The difference in intelligibility between 13,600 ft. and sea level is not statistically reliable;

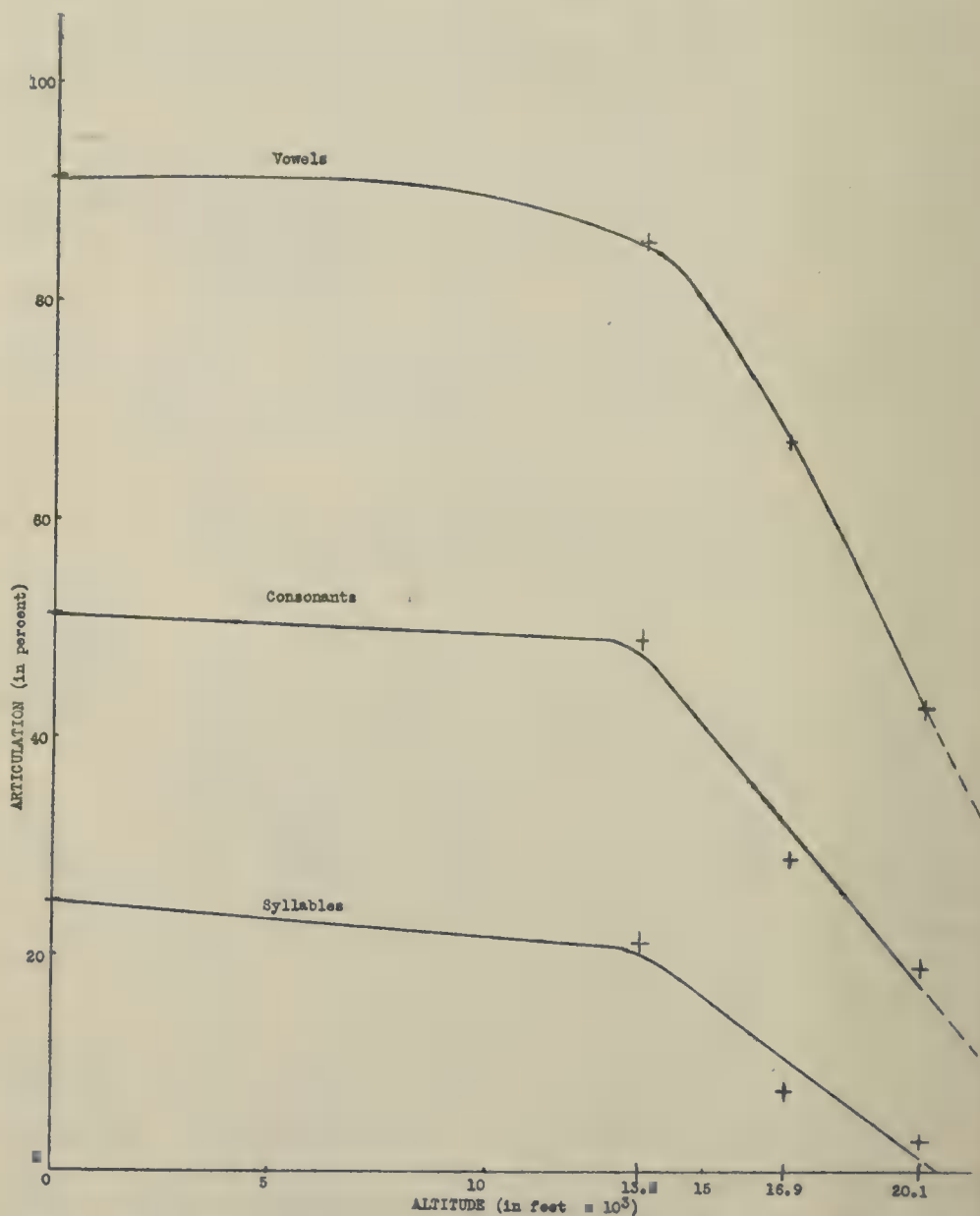


FIG. 1. The effect of altitude (anoxia) on vowels, consonants, and standard syllables. For the order of difficulty of intelligibility used in this experiment, a critical value appears between 13,600 ft. and 16,900 ft.

however, the differences in intelligibility between 16,900 ft. and sea level and between 20,100 ft. and sea level are very reliable. The difference between 16,900 ft. and 13,600 ft. is also reliable, and the difference between 20,100 ft. and 16,900 ft. is almost reliable. The degree of reliability of these differences is indicated by the column of P values, which indicate the probability that any obtained difference between means is due to chance. In general, the differences may be regarded as reliable for values of P equal to .01 or less. Even for $P = .02$ a fair reliability is indicated.

We get some light on the question of whether or not decreased articulation values at altitude are attributable to extra-auditory factors from a consideration of the next to the last row in Table 3. Here tests 2 and 11 presented at the high intensity level at 20,100 ft. are compared with the main series of tests (3-10) which were presented at the low sound intensity at the same altitude. The performance, as indicated by the highly reliable difference between the mean syllable articulation values, is vastly superior at the higher sound level. This indicates that the low articulation values obtained at 20,100 ft. in the main test series at the low sound intensity were not spurious; they were not merely the result of the subjects' inability to manipulate their pencils or to locate the proper items on the check lists, but were due to difficulties in auditory perception. That there was a small though not statistically reliable decrement in intelligibility at 20,100 ft. even at the high sound level is indicated in the last row of the same table. These results have already been discussed.

Individual differences in the effect of altitude on the ability to perceive speech sounds are shown in Table 4. The ratio of syllable articulation at sea level to syllable articulation at each of the three different altitudes is presented for each subject. A ratio greater than one indicates a decrease of intelligibility at altitude; a ratio less than one indicates an increase in intelligibility at altitude. It will be seen that even at 13,600 ft. 67% of the subjects show some loss of efficiency in perception, at 16,900 ft. 92% are affected, and at 20,100 ft. 100% are affected. Centering our attention on the ratios for the two higher altitudes, where the mean decrements have been shown to be statistically reliable, we observe a wide range in individual reactions to anoxia. In the third column the ratios vary from 1.00 to 30.0, indicating that syllable articulation for the most tolerant individual was the same at 16,900 ft. as at sea level, whereas the most susceptible individual had a syllable articulation at this altitude which was only $1/30$ of what it was at sea level. In the last column of this table a similar wide range of ratios appears. In general, those subjects who were least affected at 16,900 ft. were least affected at 20,100 ft. and those who were most affected at the lower altitude were most affected at the higher altitude. As a check on the consistency of susceptibility

Table 4
Individual Differences in Reaction to Altitude: Ratios of Syllable Articulation
Values at Sea Level to Values at Altitude

Subject	S.L./13,600 Ft.	S.L./16,900 Ft.	S.L./20,100 Ft.
1	1.2	2.3	42.0
2	1.8	1.7	14.8
3	3.3	1.0	1.7
4	2.7	10.3	8.0
5	1.1	18.0	18.0
6	0.8	10.3	18.0
7	0.5	5.7	—*
8	2.6	21.0	—*
9	0.3	30.0	—*
10	1.2	1.1	1.6
11	0.6	2.5	15.0
12	1.4	2.8	4.6
Mean	1.5	8.9	13.7*
Per Cent of Cases Showing Altitude Effect	67%	92%	100%

* Subjects 7, 8, and 9 obtained a syllable articulation value of 0 at 20,100 ft. The mean is based on the other 9 cases only.

to anoxia, as indicated by these ratios, the rank difference correlation between the values in the last two columns was calculated. It was found to be 0.70, with a probable error of 0.10.¹¹ This indicates that the rank order of the ratios S.L./16,900 ft. does not differ widely from that for the S.L./20,100 ft. ratios. This not only implies a fair degree of consistency in individual reactions to altitude, but it is a rough check on the reliability of the testing procedure.

Discussion

It should again be emphasized that the marked and reliable decrements in speech intelligibility observed at the two higher altitudes investigated in this study are not absolute values. They are definitely a function of the initial difficulty of the tests, as determined by the original speech production, by the fidelity of the recording and reproducing systems, and by the sound level employed. A question may be raised as to the significance of the results of the main series of tests which were intentionally run at a low sound level. The answer to this question is given in our earlier paper which reproduced a graph from the report of

¹¹ The rank difference correlation between the *S* values themselves for 16,900 ft. and 20,100 ft. is 0.77, with a P.E. of 0.08.

Fletcher and Steinberg (2) showing the relationship between syllable articulation and discrete word intelligibility. This indicates that when syllable articulation drops much below a point in the neighborhood of 30%, discrete word intelligibility falls precipitately and speech rapidly becomes unintelligible. It is further pointed out in our earlier report that no communication system permits perfect intelligibility under conditions of flight, because of cockpit noise, static, vibration, low pressure, poor speech production, and the special stresses of storm and combat. Hence, the additional factor of anoxia may reduce articulation, and consequently speech intelligibility, below the critical point.

Sixty-seven per cent of the subjects were affected at 13,600 ft. If such a decrement in speech intelligibility occurs in short flights of $1\frac{1}{4}$ hrs., one has additional evidence supporting the need for using oxygen on extended flights which, though generally not above 10–12,000 ft., may be of 10 to 12 hours duration. One hopes that at the higher altitudes oxygen equipment will be employed and employed effectively. With the practical introduction of pressurization the importance of anoxia as a problem for crews using conventional oxygen equipment and flying above 35,000 ft. should cease to exist.

Summary and Conclusions

Under standard conditions speech intelligibility is shown to decrease with increasing altitude. If the original sound level is low the decrement in articulation is large, discrete word intelligibility falls precipitately and speech rapidly becomes unintelligible. The effects of anoxia are minimal if sea level articulation is high. (As was pointed out in our earlier paper (3), this will be the case when: the sound level is high; a superior communication system is employed; personnel are trained in speech production and in listening under noisy and distracting conditions; and when "filling in" is made easy by familiarity with the type of message transmitted.)

The fact that 67% of the subjects (8 of 12) showed a decrement in performance at 13,600 ft. suggests the importance of using oxygen equipment on long flights even at low altitudes.

Received January 8, 1945.

References

1. Fisher, R. A. *Statistical methods for research workers*. London: Oliver and Boyd, 1938.
2. Fletcher, H., and Steinberg, J. C. *Articulation testing methods*, Bell Telephone Laboratories, Reprint B-436, November 1929, pp. 41, 42, 46.
3. Seitz, C. P., and Smith, G. M. Auditory sensitivity under conditions of anoxia; ■ study of speech intelligibility, *J. Aeronautical Sciences*, 1942, 9, 478–480.

Book Reviews

Peatman, J. G., and Hallonquist, Tore. *The patterning of listener attitudes toward radio broadcasts.* (Applied Psychology Monographs, No. 4.) Stanford University: Stanford University Press, 1945. Pp. 58. \$1.00.

The principal purpose of this monograph is to illustrate the application of cluster analysis to Program Analyzer data. Each of two radio programs was divided into program units. Tetrachoric correlation coefficients (computed with the aid of Thurstone tables) were used to express the intercorrelation among units. Cluster analysis (after Tryon) was used to find whether the program units could be considered as members of one basic cluster or whether there were several factors in operation. The theme of the monograph appears to be that having a radio program with one basic audience mood (a one-factor show) is desirable and that this method provides a way of finding out whether a program does appeal mainly to one basic audience mood. The use of cluster analysis is proposed merely as a supplement to the usual treatment of Program Analyzer data.

Three estimates of reliability were made. Two of them were estimates of the reliability with which the show as a whole was measured. The third estimate was based upon the intercorrelation of program units within each cluster. The reliability coefficients for the clusters all were above .90 with most of them over .95.

The theme of the monograph, that a one-factor show is desirable from the standpoint of acceptance, is supported more by logic and experience than by the data presented. Results for only two programs were given. The one-cluster show had a much higher Program Analyzer rating than the five-cluster program. However, this result would hardly justify acceptance of the hypothesis.

Nevertheless, the hypothesis appears reasonable, and it is likely that the writers' proposal of this hypothesis was based on considerably more evidence than actually has been presented in this monograph. Anyway, no claim is made that having one unified pattern is the only important consideration.

The methodological questions raised directly or by implication are as interesting as the practical possibilities. These questions go beyond what might be called test-reliability. They would include: (1) the extent to which the responses in the Program Analyzer laboratory situation indicate the reactions of people listening in their homes—a problem which the

writers fully recognize; (2) the whole question of the extent to which a relatively small group of respondents, perhaps representing a limited geographic area and possibly subject to selective errors resulting from voluntary cooperation, can reflect the reactions of listeners in general; and (3) the accuracy of estimates of reliability made with the use of the Spearman-Brown formula for data of this type.

Some readers may feel that the statistical aspects have not been covered thoroughly. They may notice that there is no discussion of the assumptions underlying the formulae used, no raising of the question whether the statistical methods are appropriate for data of this type. They may expect, but fail to find, an evaluation of possible methods of factor analysis.

However, these questions could not have been covered thoroughly without deviating substantially from the principal theme. They are very important, but they are broader problems that probably should be covered separately.

The reviewer hopes that this monograph reports only the beginning of what should turn out to be a significant research program, including: extension of the method to additional problems, empirical tests of the hypothesis that a one-factor show is desirable from the standpoint of acceptance, and investigations of the statistical and methodological problems previously suggested. Such a research program surely is off to a good start.

Alfred C. Welch

*Knox Reeves Advertising, Inc.,
Minneapolis, Minnesota*

Inbau, Fred E. *Lie detection and criminal interrogation*. Baltimore: The Williams and Wilkins Company, 1942. Pp. viii + 142. \$3.00.*

This book is intended primarily as a practical manual for criminal investigators. Its secondary objective is the stimulation of interest and research in the arts of lie detection and criminal interrogation.

The first section describes the technique of scientific lie detection using the Keeler Polygraph to detect the physiological concomitants of emotional reactions accompanying deception. The history, legal status, validity, and practical methodology of the technique are discussed. The effects of nervousness and fear, unresponsiveness, physiological and psychological abnormalities, and the like receive detailed treatment. Ingenious methods for meeting such contingencies are described.

According to Inbau, an accurate deception diagnosis can be made by a competent examiner in about 70% of the cases; while the indications

* The views expressed herein are the personal views of the writer and do not reflect the views of the United States Navy or the naval service in general.

may be too ambiguous or indefinite in about 20% of the cases to justify a definite statement. *And in about 10% of the cases the best examiner may be wrong.* Inbau takes this latter figure to represent the actual margin of error of the technique. He states that the chief source of error is in the failure to detect deception rather than in the misinterpretation of innocent truthfulness.

The second section describes practical techniques for criminal interrogation without lie detector aid. Suspects are classified according to certainty of guilt and probable degree of emotional conflict regarding the offense. Strategems appropriate to each class are presented.

As a practical manual this book is outstanding, but the lack of theoretical emphasis and sophistication may impair its research stimulating value. Inbau states, "The lie detector technique is predicated upon the theory that deception criteria appear in a record because of the emotional disturbances resulting from the subject's consciousness of lying and his fear of detection." But he also implicitly recognizes the importance of emotional conditioning or association, for instance, in his discussion of the possible after-effects of "third degree" examinations on the records of innocent, truthful persons. After such an experience, a truthful denial of guilt may be accompanied by emotional responses and polygraph records like those characteristic of deception. Moreover, Inbau states that persons who have successfully rationalized their crimes or who have been relieved of their emotional conflicts regarding an offense by confession to a clergyman may show few or no deception indices. And these indices of deception are absent while the criminal consciously lies concerning his offense presumably because of his fear of detection. Though "consciousness of lying" and "fear of detection" are important sources of emotional reactions, they may have been overemphasized in previous theoretical discussions of this subject. They seem to lack sufficient generality to be entirely suitable bases for either explicit or implicit theories of lie detection. Perhaps a theory formulated in terms of emotional conditioning or association might have both greater generality and greater stimulus value for psychologically meaningful experimentation.

Howard F. Hunt

United States Naval Reserve

De Silva, Harry R. *Why do we have automobile accidents?* New York: John Wiley & Sons, Inc., 1942. Pp. ix + 394. \$4.00.

This book deals with some phases of the important problem of highway mishaps with emphasis on the psychological aspects of accident causation. An analysis is made of the contributing factors, such as exposure to accidents, speed, skill at the wheel and attitudes toward

safety. Data are given on the types of drivers behind the wheel and exposure-in-travel hazards. Characteristics of the driver, such as age, reaction time, experience and skill in general, are treated in relation to speed. It is stated that fifty miles an hour on the best of highways is fast enough for the average driver.

The author presents his own adaptation of tests for drivers which includes both road and laboratory procedures. Considerable attention is given to vision in relation to safe driving. The treatment in this section is somewhat stinted and overlooks the work done by many other investigators.

One section of the book is devoted to sociological aspects of accident susceptibility. Skill alone is said to be only a part of safe driving. Some evidence is given which indicates a higher accident ratio in the lower socio-economic groups. The author contends that illiteracy and foreign birth are added susceptibilities to accident, but no statistical analysis is offered to show that these factors operate independently of other conditions and attitudes. Consideration is given of the effects of intoxication, fatigue, ill health and related factors but the treatment is limited in scope.

A very satisfactory survey is made of methods used in collecting accident information, types of accidents, reporting systems, interpretation of data on accidents and criteria used in safety contests. Financial responsibility laws and compulsory insurance are discussed in a general way.

The last part of the book deals with various related problems, including pedestrian traffic, comparative figures on urban and rural accidents, enforcement, outstanding characteristics of drivers, factors concerning road conditions in relation to accidents and re-examination procedures of drivers. A research program on highway safety is presented at the conclusion of the book.

In all there are 374 pages and a bibliography of 300 listed references. A reasonable amount of tabular, graphic, and illustrative material is included. The index is rather incomplete, possible due in part to the wide coverage of topics within the volume. Although it might be much better organized, the book serves as an introduction to this very much neglected area of study, on the part of academic writers, and should be read by everyone interested in traffic problems and automobile accident prevention. It should, however, be read somewhat critically as the presentation is sketchy in places. The author describes his treatment in the preface as an attempt, "to offer the reader a bird's-eye view of the field."

A. R. Lauer

*Iowa State College,
Ames, Iowa*

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to
Donald G. Paterson, Editor, Department of Psychology, University
of Minnesota, Minneapolis 14, Minnesota

- Job safety training manual.* Kenneth L. Faist, National Foremen's Institute, Inc., Deep River, Conn., 1944. Pp. 52. \$5.00.
- Training for supervision in industry.* George H. Fern. New York: McGraw-Hill Book Co., Inc., 1945. Pp. 188. \$2.00.
- Understanding labor.* Bernard H. Fitzpatrick. New York: McGraw-Hill Book Co., Inc., 1945. Pp. 179. \$2.00.
- Conference leadership in business and industry.* Earle S. Hannaford. New York: McGraw-Hill Book Co., Inc., 1945. Pp. 289. \$3.00.
- Top-management planning.* Edward H. Hempel. New York: Harper & Brothers, 1945. Pp. 414. \$4.50.
- How to get the job you want!* John W. Herdegen. New York: Essential Books, Inc., 1945. Pp. 92. \$1.00.
- Sell America into jobs.* William E. Holler. Motor City Publishing Co., General Motors Bldg., Detroit 2, Mich., 1945. Pp. 128. \$1.50.
- New goals for old age.* George Lawton. New York: Columbia University Press, 1943. Pp. 210. \$2.75.
- Human leadership in industry: The challenge of tomorrow.* Sam A. Lew-iso-hn. New York: Harper & Brothers, 1945. Pp. 112. \$2.00.
- The social problems of an industrial civilization.* Elton Mayo. Division of Research, Harvard Business School, Soldiers Field, Boston 63, Mass., 1945. Pp. 150. \$2.50.
- Counseling service for industrial workers.* Mary Palevsky. New York: Family Welfare Association of America, 1945. Pp. 51. \$.60.
- Diagnostic psychological testing, Vols. 1 & 2.* David Rapaport. Chicago: The Year Book Publishers, Inc., 1945. Pp. 573. \$6.50 per vol.
- Counseling with returned servicemen.* C. R. Rogers and J. L. Wallen. New York: McGraw-Hill Book Co., Inc., 1945. Pp. 165. \$1.60.
- Sex and the social order.* G. H. Seward. New York: McGraw-Hill Book Co., Inc., 1945. Pp. 286. \$3.50.
- Where do people take their troubles?* Lee R. Steiner. Boston: Houghton Mifflin Co., 1945. Pp. 263. \$3.00.
- Productive thinking.* Max Wertheimer. New York: Harper & Bros., 1946. Pp. 224. \$3.00.

- The relation between illumination and visual efficiency—the effect of brightness contrast.* H. C. Weston, Report No. 87, Industrial Health Research Board, Medical Research Council, London, England, 1945. H. M. Stationery Office, York House, Kingsway, London, W. C. 2. Pp. 35. 9d.
- Procedures in evaluating a guidance program.* Frances Morgan Wilson. New York: Bureau of Publications, Teachers College, Columbia University, 1945. Pp. 210. \$2.60.
- A study of women on war work in four factories.* Industrial Health Research Board Report No. 88 of the Medical Research Council. S. Wyatt *et al.* H. M. Stationery Office, York House, Kingsway, London, W. C. 2, England, 1945. Pp. 44. 9d.
- Job descriptions for office occupations.* Washington 25, D. C.: Superintendent of Documents, U. S. Government Printing Office. 1945. Pp. 204. \$1.25. Division of Occupational Analysis, War Manpower Commission.
- Occupational data for counselors:* A handbook of census information selected for use in guidance. Bulletin No. 817, Bureau of Labor Statistics, U. S. Department of Labor, 1945. Washington 25, D. C.: Superintendent of Documents, U. S. Government Printing Office, Pp. 36. \$.10.
- Racial aspects of reconversion:* A memorandum prepared for the President of the United States. New York: National Urban League, 1945. Pp. 29. Gratis. (Cost of postage charged for bulk orders.)
- The placement interview.* Civilian Personnel Pamphlet No. 15. War Department, Washington 25, D. C. Pp. 17.
- Training and research in industrial relations.* Proceedings of May 1945 Conference. Bulletin 1. Industrial Relations Center, University of Minnesota. October, 1945. Pp. 1-60. Gratis.

Journal of Applied Psychology

EDITED BY: DONALD G. PATERSON, UNIVERSITY OF MINNESOTA

Consulting Editors

UL S. ACHILLES, *Psychological Corporation*; WALTER V. BINGHAM, *A.G.O., War Department*;
 HAROLD E. BURTT, *Ohio State University*; ARTHUR I. GATES, *T. C. Columbia University*;
 JOHN G. JENKINS, *University of Maryland*; IRVING LORGE, *T. C. Columbia University*;
 VINN MCNEMAR, *Stanford University*; WILLARD C. OLSON, *University of Michigan*;
 JAMES P. PORTER, *Swarthmore, Pennsylvania*; EDWARD K. STRONG, JR., *Stanford University*;
 MORRIS S. VITELES, *University of Pennsylvania*; JOSEPH ZUBIN, *N. Y. Psychiatric Institute*.

Table of Contents

<i>Output Rates Among Butter Wrappers: I. Work Curves and Their Stability:</i>	
	HAROLD F. ROTHE 199
<i>Management's Reactions to Employee Opinion Polls:</i>	ROBERT N. McMURRY . 212
<i>Signed Versus Unsigned Personal Questionnaires:</i>	ROBERT P. FISCHER 220
<i>Life of College Graduation and Success in Adult Life:</i>	SIDNEY L. PRESSEY . . 226
<i>Predicting Success in a School of Nursing:</i>	A. Q. SARTAIN 234
<i>Teachers College Students and the Minnesota Multiphasic Personality Inventory:</i>	ORPHA MAUST LOUGH 241
<i>The Occupational Adjustment Characteristics of a Group of Sexually Promiscuous and Venereally Infected Females:</i>	ROBERT D. WEITZ 248
<i>The Effect of Prolonged Mild Anoxia on Speech Intelligibility:</i>	
	G. MILTON SMITH 255
<i>Studies in International Morse Code: V. The Effect of the "Phonetic Equivalent":</i>	F. S. KELLER, I. J. CHRISTO, AND W. N. SCHOENFELD 265
<i>Validity of the Hunt-Minnesota Test for Organic Brain Damage:</i>	
	RACHEL F. MALAMUD 271
<i>The Hunt-Minnesota Test for Organic Brain Damage in Cases of Functional Depression:</i>	PAUL E. MEEHL AND MARY JEFFERY 276
<i>Book Reviews</i> 288
<i>New Books, Monographs, and Pamphlets</i> 295

Published Bi-monthly by The American Psychological Association, Inc.

Prince and Lemon Sts., Lancaster, Pa., and

Massachusetts and Nebraska Aves., NW, Washington 16, D. C.

Registered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the Act of March 3, 1879
 Copyright, 1946, by The American Psychological Association, Inc.

Journal of Applied Psychology

Vol. 30, No. 3

June, 1946

Output Rates Among Butter Wrappers: I. Work Curves and Their Stability *

Harold F. Rothe

Stevenson, Jordan and Harrison, Inc., Chicago, Illinois

The importance of production data in industrial situations has long been recognized. This is true whether these data are used in the analysis of engineering problems or of behavioral problems. As Burt has written, production is the most obvious criterion to use in validating a selective personnel test (3, 173). Muscio has stressed the use of work curves in studying problems of "industrial fatigue" (8). Production data, in one form or another, have also been used in analyzing the effects of variation in illumination, atmosphere, wage systems, work methods, training methods, and so forth.

Despite this rather wide usage of output data, very little research has been published on the various problems of this criterion. For many industrial plants little or nothing is known about the pattern, distribution, and stability of rates of output of various employees under various environmental conditions. Because so little is known about these three phenomena it is impossible to generalize about them to any great extent, or to make predictions about any given new situation. This is a serious obstacle in the development of a scientific industrial management and a scientific industrial psychology.

This paper reports an investigation that was made of these problems in a particular plant for a specific type of operation. The data were collected in an industrial situation and they were analyzed according to the methods of experimental psychology. The findings in regard to the patterning of rates of output and its stability are presented here. The

* This study was made in partial fulfillment of the requirements for the degree Ph.D. in psychology at the University of Minnesota. The writer wishes to thank his co-chairmen, Professors Donald G. Paterson and Miles A. Tinker, for their many helpful suggestions. He is also grateful to Mr. John Brandt, President, and the employees of the Minneapolis plant, Land O'Lakes Creameries, Inc., whose cooperation made possible this investigation.

findings in regard to the distribution of output data will be reported in a second paper.

Pertinent Literature

Before discussing briefly the pertinent publications in these areas it might be well to define the terms "pattern" and "stability" as they are used here. The words "pattern" or "patterning" of output data indicate the distribution of production rates in time so that when they are presented graphically the result is a so-called work curve or production curve. The term "stability" as used here is analogous to the term "reliability" as that is commonly used in psychological work. Thus if an operator shows perfectly identical work curves for two different days the curves are said to be "stable" (within the limitations of the two days).

There are two rather general concepts of work curves. Burt's so-called "typical daily work curve" is so frequently illustrated in textbooks that the impression is often left with the reader that this is the only shape in which work curves are found (2, 154). Indeed, approximations of this curve have been found as by Goldmark and Hopkins (9, 429), and by Polatov (10). This work curve (sometimes also called "fatigue" curve) has been found most often in heavy operations. It is found when some kind of production data for many operators are lumped together and there is but little published evidence of this "typical" curve for any one operator on any one day.

The concept of a typical "monotony" curve has been suggested by Wyatt (14). This curve is most frequently associated with light repetitive operations and reported feelings of monotony, as by Marsh (6) and by Wyatt, Frost, and Stock (15). It, too, is based primarily upon group data and there is very little indication that it is found for any one operator on one day.

The Hawthorne investigators reported a few daily work curves for some of their relay assembly test room operators and these curves did not appear to fall readily into either of the above two classifications (11, 121). They appeared as plus and minus variations about a straight line parallel to the abscissa. Composite curves of this straight line nature were also found by Wyatt, Frost and Stock (15).

The stability of industrial work curves has been investigated very slightly and apparently by inspectional methods only. Daily work curves specific to the days on which they were made by individual operators were reported by Vernon (13) and by operators singly and in a group whose data were considered together by Kunst (4).

The questions of whether work curves are specific to the type of operation and to the methods of payment were investigated by some of

the English students. Burnett found average work curves to vary separately with method of payment (piece rate and time rate) and also with day of the week as well as to co-vary with these (1). Wyatt, Frost, and Stock found average curves to vary with method of payment but not with type of operations (15).

In summary: (1) individual and group curves may take any of a number of shapes; (2) there are few published data regarding the shapes of the work curves of individuals operators on single days; (3) work curves, for both individual and groups, *may* vary "capriciously" from day to day, *may* vary with methods of payment, and *may* vary more with variations in payment than with variations in the nature of the work; and (4) in the light of the above, one is treading upon extremely hazardous ground when he attempts to predict anything about the pattern or stability of work curves for any individual or group in a new situation.

The Present Problems

The relative paucity of industrial data indicated above presents a severe problem to the industrial psychologist studying the effects of rest periods, music, etc. in factories and offices. Although he may wish to use work curves as a validity criterion of his work, he often cannot know how much data to collect to insure a stable criterion. Frequently, too, his opportunities for collecting data will be seriously limited. It is desirable, therefore, to find further facts regarding the pattern and stability of output data in order to answer some of his questions.

In the present study it was possible to obtain data on some industrial operators performing a light, repetitive, manual operation. The problems, then, were to determine the patterning of these output data and the stability of this patterning for the operators individually and as a group.

Conditions and Methods of the Investigation

In an experimental investigation it is possible to control many variables such as time, place, work method, and others. In an industrial investigation this is usually impossible and the investigator can only observe and describe the conditions. The latter was true of this study.

The study was made in the print room of a Minneapolis creamery. Data were collected during the "regular" working hours from Monday, March 15, 1943, through Monday, March 29, 1943. The hours of work were from 7:30 A.M. to 4:30 P.M. on Monday through Friday, and from 7:30 A.M. to 12:00 noon on Saturday. There were daily rest periods from 9:30 A.M. to 9:40 A.M. and from 2:30 P.M. to 2:40 P.M. The lunch period was from 11:30 A.M. to 12:00 noon, daily. These hours represent some overtime because of war conditions. The operators did not know, on any day, how late they would

work that afternoon until shortly before quitting time. They actually worked the above hours on each day during the study. Weekly pay was by time plus overtime.

The temperature of the room was measured several times during the study with an inexpensive thermometer. In general it varied between 66 degrees and 74 degrees Fahrenheit, and it rose slowly but steadily during each day. The humidity was not measured but was necessarily somewhat high because of the nature of some of the operations in the print room. The illumination was furnished by means of windows and skylights and also artificially. The foot-candle illumination at the work-table was taken several times and was found to range from 7 to 13 f.c. This was more than sufficient for the type of work performed (12, p. 19).

During the two weeks of the study seven different operations were performed. Of these, one was by far the most common of the seven. This operation consisted of wrapping, by hand (to supplement the wrapping machines in the room) quarter-pound blocks of butter. These were taken off a moving belt, wrapped in single vegetable parchment wrapping papers, and replaced on a higher belt. Close observation revealed that there were so many slight variations in techniques among the operators that a detailed motion analysis was unfeasible for the present purpose.

The butter was wrapped on a table, the top of which was 32 inches above the floor. The operators sat in adjustable chairs along both sides of the table. These chairs were movable and no attempt was made to have each operator sit in the same position each day although they tended to do this.

Cenco counters were placed on the table before each operator. Each counter had a small arm which the operator depressed whenever she placed one pound of butter (wrapped in four quarters) on the upper belt. An assistant went around the table, in the same order, once each fifteen minutes, and recorded the counter readings. On occasion other readings were taken. The investigator sat at one end of the table where he could see all operators and record all activity other than butter-wrapping. Thus a measure was obtained of any inactive time within a fifteen minute period. The data were adjusted for such periods by dividing the output for the period by the number of active minutes (within five seconds) and multiplying by fifteen. This was done for all periods of inactivity under five minutes in length in any fifteen minute period. This involved the assumption that the work within any fifteen minute period was spread evenly over that interval for each operator. The use of the counters required the assumption that the operators' working habits were not seriously impaired. This was probably partially true because the management had previously, on occasion, used these counters for a somewhat similar purpose. The data for all fifteen minute periods in which the operators were inactive for more than five minutes were omitted from the analysis.

The operators who were observed in this study were regular employees of the creamery. All were experienced at hand wrapping. All were women, ranging in age from 18 to 39 years, in education from 8 to 12 years, and in length of service from one month to eighteen years. All except the one with one month's service were Union members. Data were collected for 16 operators. Many of these were incomplete because of necessary job shifting, i.e., changes in assignments. The above personnel facts, and all of the production data discussed in these two papers, were based upon the *eight* operators for whom complete analyses were made.

It is unnecessary to dwell upon the variables that were not controlled in this investigation. This is properly spoken of as an investigation and not an experiment. The following uncontrolled sources of variation may be briefly noted: relatively coarse timing, non-mechanical production counting, shifting numbers and positions of operators along work-table, presence of investigators

and other visitors to the print room, rate of flow of butter along belts, hardness of butter, atmosphere, illumination, humidity, methods of wrapping, effects of shifting operations, age, experience, and private lives of operators, amount of talking, and necessity for substitute research assistants in last few days of the study. The influence of any and all of these factors can only be surmised.

The data used in this study for an analysis of the pattern, distribution, and stability of the output of industrial workers under "industrial conditions" were those collected for eight operators only. The data collected on the other operators were considered too incomplete to permit of extensive analysis. These data, as mentioned above, were also "adjusted" to furnish "complete" fifteen-minute intervals of production in those periods when the operators were away from the table or otherwise not busy.

Results

Daily work curves were constructed for each operator for each day, in the following manner. Units of time, both hours and days, were indicated on the abscissa, and rates of production expressed in terms of the number of pounds wrapped within a fifteen minute period, were indicated on the ordinate. Differences in operations were indicated by a code; rest periods and lunch periods by ordinates erected from the base line and one curve for each of the entire thirteen days was made for each operator.¹

Inspection of these curves by the writer and by the members of his Ph.D. thesis committee of the Graduate School of the University of Minnesota, indicated that they took many different, and probably no characteristic, shape. In some instances they resembled "fatigue" curves; others resembled "monotony" curves; and the greater majority of them were more or less "straight-line" curves. There was nothing in the data to permit any prediction as to the kind of curve that might be expected for any individual on any day, with the exception of the facts discussed below in the correlational analysis. The curves for each operator varied among themselves and did not appear, by inspection, to be consistent from day to day.

A so-called group daily curve was constructed for each day in the following manner. For every fifteen minute period throughout the study, the available production rates of all operators were listed. The median for each of these periods was used as the output rate for that period on this group curve. Thus each point on it is the median of eight individual readings, except for those periods when some of the operators were not working. At no time were less than three readings used in establishing this median.

¹ These original curves and all raw data have been placed on file with Dr. Miles A. Tinker in the Department of Psychology at the University of Minnesota. Photostatic copies of the curves are in the writer's thesis on file in the University of Minnesota Library.

This group daily work curve likewise failed to show any consistent common patterning, by inspection, with these two exceptions: (1) the straight-line curve was again the most common, although "fatigue" and "monotony" curves were also in evidence; and (2) rest periods were frequently followed by a lowered production and lunch periods by a higher production.

A "daily trend line" was established for each operator. Each operator's median output rate for every fifteen minute interval, regardless of days, was determined, and these medians were connected to make an "average" or "trend" line that was assumed to be independent of any influence that could be attributed to particular days of the week. These trend lines were also analyzed by inspection. Those for two operators resembled the "typical work curves" or "fatigue" curves. Two others showed what might be considered "monotony" curves, and the other four operators had "mixed" trend lines. One characteristic of all trend lines was that the lowest point for every operator occurred in the first fifteen minute period of the day, and the lowest prolonged period (plateau) occurred in the late afternoon. This latter might be considered evidence of "fatigue," or better, could be defined as fatigue.

Group trend lines were also established. Two of these were made, using two techniques. The first was to "collapse" all of the group daily work curves on to one curve, thus eliminating the effects of variations among days. The second method was to locate the points from each of the individual trend lines onto one graph and to connect the medians of these for each fifteen minute interval. These two group trend lines showed the same general characteristics, as might be expected. They both began low and rose steadily in the first morning work-spell, levelled off in the second morning work spell, began at a lower level after lunch but rose quickly until they reached a peak for the day just before the afternoon rest period, and then showed a "dip," resembling a "monotony" curve, in the late afternoon spell while remaining at a generally lower level in this last period.

In summary, the individual trend lines varied from operator to operator, taking several general shapes. The individual and also the group trend lines showed a few common phenomena such as an early morning warming-up period, a tendency toward a general decrease during the day, and an apparently relatively inefficient spacing of the rest periods.

In addition to the inspectional analyses described above, the work curves were analyzed by correlational techniques. The method was, in general, to correlate one curve with another curve, obtaining Pearson r 's. The paired variates were the two readings for any given fifteen minute period.

Although no correlations were computed where there were not at least 18 paired variates, most correlations involved about 30 sets of readings. In order to minimize the effects of these small samples, the practise was to determine a large number of correlations dealing with each question (as far as was possible) and then to consider the distribution of these coefficients, rather than to attach much significance to any one of them. Further, in determining these correlations the data for five days only were utilized. On these days all of the eight operators performed the same operation (wrapping quarter pound blocks of butter) all day long. These days were Monday 2, (the second Monday of the

Table 1

Distribution of Correlation Coefficients between Work Curves of Different Days
for Same Operators

Range of r	Frequency
.60 to .69	2
.50 to .59	1
.40 to .49	5
.30 to .39	5
.20 to .29	5
.10 to .19	8
.00 to .09	7*
-.01 to -.10	12
-.11 to -.20	7
-.21 to -.30	4
-.31 to -.40	4
-.41 to -.50	1

* Indicates location of median of distribution.

investigation), Tuesday, Wednesday, and Thursday (also all of the second week), and Monday 3.

The first correlational problem attacked was that of the stability of an individual operator's daily work curve from day to day. This is analogous to the problem of test-retest reliability. To answer this problem, each operator's work curve for one day was correlated with her work curve for every other day. There were ten such pairs of days and eight operators. On a few occasions some operators missed some of these days and hence only 61 (rather than 80) correlations were obtained. The distribution of these correlations is shown in Table 1.

Thus, using the median coefficient of the distribution in Table 1, the inter-day correlation between the work curves for any one operator was negligible (approximately .05).

To ascertain the possible presence of any sort of group social interacting (i.e., talking, common reaction to events or days of the week, etc.) a series of coefficients was determined for the work curves of individual operators on common days. That is, the correlation between Operators 1 and 2, both on Tuesday, etc. It did not appear necessary to correlate all possible combinations and a sample of 35 combinations was selected with the aid of tables of random numbers (5, 262 ff.). The distribution of the obtained coefficients is shown in Table 2, where the median is shown to be about .27.

From Tables 1 and 2 it may be concluded that the work curves for the various operators showed a slight tendency to take the same shape on

Table 2
Distribution of Correlation Coefficients between Work Curves of Different Operators on the Same Days

Range of r	Frequency
.70 to .79	1
.60 to .69	0
.50 to .59	5
.40 to .49	7
.30 to .39	3
.20 to .29	8*
.10 to .19	4
.00 to .09	1
-.01 to -.10	1
-.11 to -.20	3
-.21 to -.30	0
-.31 to -.40	1
-.41 to -.50	1

* Indicates location of median of distribution.

any one day while the shape of any one operator's work curves bore no relation to each other from day to day. This finding, here by correlational techniques, is in accordance with the conclusions of the studies described by Roethlisberger and Dickson (11), Wyatt, Frost, and Stock (15), and Mayo and Lombard (7).

In order to test further the inter-day relations among the work curves for each operator, another series of correlation coefficients was obtained in which the curve for one operator on one day was correlated with the curve for a different operator on another day. Thirty-five such pairs of curves were selected, again with the aid of tables of random numbers. The distribution of these is shown in Table 3.

The median of the distribution in Table 3 is .05. Thus, as can be

Table 3
Distribution of Correlation Coefficients between Randomly Paired
Individual Work Curves

Range of r	Frequency
.60 to .69	1
.50 to .59	2
.40 to .49	3
.30 to .39	4
.20 to .29	1
.10 to .19	6
.00 to .09	5*
-.01 to -.10	7
-.11 to -.20	2
-.21 to -.30	2
-.31 to -.40	1
-.41 to -.50	1

* Indicates location of median of distribution.

seen from Tables 1, 2, and 3, an operator's work curve for any one day bears little or no relation to her own curve for any other day or to the curve of any other operator for any other day. But when the work curves for different operators on any one day are considered, there is some relationship.

Using these same correlational techniques the group daily work curves were also analyzed to determine if they were stable from day to day. It would be anticipated from the above that this might be true. The matrix of these coefficients is presented in Table 4.

Table 4
Correlation Coefficients between Group Work Curves for Different Days

	Tuesday	Wednesday	Thursday	Monday 3
Monday 2	.22	.34	.31	.08
Tuesday		.26	.53	-.09
Wednesday			.50	.34
Thursday				-.22

These coefficients range from .53 to -.22 and the median is about .30. Thus the group as a whole tended to show the same work pattern from day to day, although the correspondence was not very high. These correlations were highest between Tuesday, Wednesday, and Thursday. It is possible that the relatively peculiar work curves for Monday 2 and Monday 3 might be functions of the week-ends or of week-ending. The

data for Saturdays have been omitted from all correlations because of the small number of available readings.

In a similar manner, the individual trend lines were also correlated, and since the group daily curves were related, it appeared probable that the trend lines which amounted to "averages" for individuals over a period of days might be related to an appreciable degree. This was found to be so, as shown in Table 5, where the range of coefficients is from .15 to .70, with the median at .51.

Table 5
Correlation Coefficients between Trend Lines for Each Operator

Operator Number	2	3	4	5	6	7	8
1	.68	.39	.62	.70	.51	.61	.15
2		.27	.59	.54	.51	.47	.33
3			.66	.57	.48	.33	.55
4				.59	.60	.63	.43
5					.47	.62	.49
6						.52	.43
7							.42

The distribution of these coefficients is shown in Table 6. This distribution is skewed as would be expected if there were actually some positive correlation.

Table 6
Distribution of Correlation Coefficients between Trend Lines for Each Operator

Range of r	Frequency
.61 to .70	7
.51 to .60	9
.41 to .50	7
.31 to .40	3
.21 to .30	1
.11 to .20	1

It appears, then, that the different operators tended to vary together in production pattern when a fairly long period of time was considered, i.e., several days. The reasons for this could not be determined from the present data, although it is probable that the group interaction or the group common reaction mentioned above may have been influential. Another possible reason for this tendency toward higher correlations among trend lines than among daily work curves is that the trend lines

are more "stable" or "reliable" than are the daily curves. It is also possible that these inter-individual trend lines are higher than are the correlations among the group curves from day to day because the trend lines are constructed from data covering several days and hence include inter-day variation as well as intra-day variation. The group daily curves include inter-individual but not inter-day variation.

All of the correlations that have been mentioned tended to be positive. Even those distributions with medians of approximately 0.00 extended further into the positive values of r than into the negative values. This suggested that the inter-correlation between the two group trend lines, both based on essentially the same data, but derived by different methods, would be high and positive. A coefficient of .87 was obtained when the two group trend lines were correlated.

To the extent that this coefficient of .87 was roughly predicted from the other coefficients, it serves partially to validate those others and to indicate that they are probably rather accurate estimates of the inter-relationships that actually prevail. And to the extent that the group trend lines were established by different methods this latter coefficient suggests that the situation they represent is a fairly stable one, and that different methods of treating sufficient data of that situation give essentially the same results.

In summary, the following correlational results may be tentatively presented: the correlation between an operator's work curve for one day vs. her curve for another day is about 0.05; the correlation between one operator's curve and the curve for any other operator on any different day is also about 0.05; the correlation between the curves for any two operators on any one day is about 0.30; the correlation between any two operators' trend lines (covering several days) is about 0.50; and the correlation between group trend lines for several days and established by different methods is about 0.90.

Summary and Conclusions

The conclusions to be drawn from this investigation must be considered tentative rather than definite, because of the serious limitations imposed by the small number of operators, the restricted nature of the operations, and the short period of time involved. Within these limitations the following general conclusions may be drawn:

1. Individual daily work curves for this, and probably other, industrial jobs involving light manual skills, may take any of many different forms and do not assume any characteristic, predictable pattern.
2. Individual daily work curves are, as the name implies, specific to the individual and also, but to a lesser extent, specific to the day. They

are correlated with each other only to the extent that the early warming-up period is a common phenomenon.

3. Individual trend lines, based on the data of several days, are more highly related among different individual industrial operators than are individual daily work curves.

4. Group trend lines, regardless of method of construction, and within the limits used in the present analysis, represent a stable phenomenon. It is to be expected that group trend lines based upon samples of different periods of time, but on the same operations, would be intercorrelated rather highly. Thus group trend lines should be used when work curves are used as criteria against which some variable is to be measured or validated since they form stable criteria.

5. The correlational technique applied to work curves is one that may well be applied more widely in future industrial research on work patterning.

6. Industrial management, in studying the effects of rest periods, music in factories and offices, illumination, etc., by analyzing the effects of these variables upon work curves, would be wise to collect data covering several different operators and several different days in order to establish a "stable" work curve of output. If this is impossible, it would be next most valuable to obtain data on one operator over several days and, as a third choice, to collect data on several operators on any one day, and thus construct a reasonably accurate estimation of the work curve for the operation under consideration. There is little or no value in collecting data for one operator on one day only.

Received June 25, 1945.

References

1. Burnett, I. An experimental investigation into repetitive work. London: *Ind. Hlth. Res. Bd.*, Rep. No. 30, 1925.
2. Burt, H. E. *Psychology and industrial efficiency*. New York: D. Appleton and Co., 1929.
3. Burt, H. E. *Principles of employment psychology*. (Rev. Ed.) New York: Harper and Brothers, 1942.
4. Kunst, E. J. Variations in work performance under normal industrial conditions. *Psychol. Bull.*, 1941, **38**, 530.
5. Lindquist, E. F. *Statistical analysis in educational research*. Boston: Houghton Mifflin, 1940.
6. Marsh, H. D. The diurnal course of efficiency. *Arch. Phil., Psychol., and Sci. Meth.*, 1906, **14**, No. 3.
7. Mayo, E., and Lombard, G. F. F. Teamwork and labor turnover in the aircraft industry of southern California. Boston: Harvard Univ. Bur. Bus. Res., Study No. 32, 1945.
8. Muscio, B. Is a fatigue test possible? *Brit. J. Psychol.*, 1921, **12**, 31-46.

9. Poffenberger, A. T. *Principles of applied psychology*. New York: Appleton-Century, 1942.
10. Polatov, W. E. Making work fascinating as the first step toward reduction of waste. *Mech. Eng.*, 1921, **43**, 731-732.
11. Roethlisberger, F. J., and Dickson, W. J. *Management and the worker*. Cambridge: Harvard Univ. Press, 1941.
12. Tinker, M. A. Illumination standards for effective and comfortable vision. *J. consult. Psychol.*, 1939, **3**, 11-20.
13. Vernon, H. M. Can laboratory experiments on output throw light on problems of industrial fatigue? *Brit. J. Psychol.*, 1924-25, **15**, 393-404.
14. Wyatt, S., Fraser, J. A., and Stock, F. G. L. The effects of monotony in work. London: *Ind. Hlth. Res. Bd.*, Rep. No. 56, 1929.
15. Wyatt, S., Frost, L., and Stock, F. G. L. Incentives in repetitive work. London: *Ind. Hlth. Res. Bd.*, Rep. No. 69, 1934.

Management's Reactions to Employee Opinion Polls

Robert N. McMurry

Robert N. McMurry & Co., Chicago, Illinois

Ordinarily the public thinks of labor disputes as having their origin in such major issues as wages and hours. This is not strictly true. While they serve to rationalize attacks on management, the true causes are the *multitude of trivial annoyances to which the employee is subjected day in and day out*. The chief reason why they are allowed to continue and to create worker hostility is the fact that management rarely knows of their existence,—hence does nothing about them.

One of the best tools for discovering the true sources of worker dissatisfaction is the so-called "Employee Opinion Poll" or the "Employee Morale Survey." Any organization which is genuinely interested in building and maintaining good employee morale should at intervals obtain a measure of its employees' attitudes by means of one of these polls. Such a poll consists of a series of questions which are asked of workers on the job to ascertain what they like and dislike about their working conditions, supervision, management policies, employee services, rates of compensation, and their attitudes toward the personalities and competence of top management. The form used is of the multiple answer type usually carrying from twenty to sixty specific questions. Each of these has four alternative answers. A typical item follows: The equipment with which I work is: a. ———Very satisfactory; b. ———In good condition; c. ———Somewhat unsatisfactory; d. ———Definitely unsatisfactory.

The employee answers the items on the form by placing a check mark before the particular one among the four responses which best represents his answer to the question. In this way a uniform measure of employee opinion is obtained and *no writing is required of the individual*. He is requested not to sign his name, or identify himself in any way, thus insuring that his responses are strictly anonymous. Furthermore, report to top management is in summary form, showing the distribution of answers to each item for each department. This further insures against any possibility of individual identification. This latter is important because, if the employee believes that his identity is not protected he may be motivated to give untruthful answers that will subject him to no risk of retaliation by supervision or others.

In practice, opinion polls are usually administered as follows: A representative of management makes a brief explanation to the employees to be polled, outlining the purpose of the project. Following this, the employees are asked each to pick a questionnaire from a pile placed on the table, so that there is no danger to an individual that his form has been keyed. After he has answered the questions, he puts the form in a ballot box on the table. After all the questionnaires have been inserted, the box is sealed and given to an outside organization for tabulation. This is done to give the employees further assurance that the identity of each is protected. Furthermore, never less than ten questionnaires are placed in any one box to avoid any additional possibility of identification.

Wherever possible, a separate ballot box is provided for supervision and for each department, shift, and supervisory unit. Where sufficient numbers are available to make it feasible, men and women have different colored questionnaires. Thus it is possible to analyze the results separately for supervision, for men and women, and for each department, shift and supervisory unit.

While some advocate the administration of these polls by mail, it is a practice to be discouraged. First, only a part of the employees will answer. Consequently, it is impossible to be certain of the extent to which the sample is truly representative. Second, a questionnaire sent to the home is likely to be filled out in consultation with other family members and often associates in the union. As a result, it is difficult to ascertain the extent to which the opinions expressed *are those of the employee*. Third, the fact that the employee is asked to mail in the form may make him question the degree to which his identity has been protected since it is easy to key individual forms. This in turn, may influence his responses by leading him to play safe by expressing few dissatisfactions.

The contribution of the employee opinion poll to the building and maintenance of employee morale is that it serves as an important medium of *communication between* the employees and management. It is necessary because, unfortunately, as companies grow larger, there is a tendency for relationship between the man on the machine or at the desk and top management to become increasingly distant. This problem does not exist in the smaller organization where direct personal contacts exist. Where this personal contact has been lost, communication between management and the worker begins to break down. On the one hand, there is a tendency for company policies, as they are transmitted *downward from*, management through the hierarchy of supervision to the worker to undergo modification or even distortion. A policy may have been sound and fair when it was approved by top management, but by

the time it is applied, it may have become so altered that it is no longer recognizable and may be far from sound and fair, yet because of this failure of communication, management is quite unaware of any change.

On the other hand, in communication *upward* from the worker to management, employee dissatisfactions and grievances are frequently blocked by members of supervision. In any organization there are always some who are desirous of creating the impression with their superiors that there is no dissatisfaction among their subordinates. Hence, they make every attempt to repress and conceal evidences of poor morale among those who report to them. Nor do they encourage their subordinates to bring their troubles to them. In consequence, top management is often not kept informed of what the employees are actually thinking or is not made aware of legitimate complaints which they may have. As a result of these failures of communication, employees are frequently subjected to unnecessary injustice and frustration. The resulting dissatisfactions, being denied outlet or redress, tend to accumulate, and create a generalized hostility toward the company and management in general. This is invariably destructive to morale.

The employee opinion poll is designed to provide one avenue of communication between the workers and top management. It provides the man on the machine or at the desk with an opportunity, free from any danger of retaliation by his superiors, to express frankly and without reservation his likes and dislikes with respect to his job, the people with whom he is associated and the company as a whole. Thus, the opinion poll brings the principal sources of employee dissatisfaction to the attention of top management, enabling it to take steps to eliminate them and to take the initiative in counteracting their effects. It also locates those sore spots in the company organization which require immediate attention so that corrective measures can be applied where they are most needed first.

Frequently, the chief sources of employee ill-will and poor morale are not such major items as wages or hours, but rather are the many petty annoyances to which workers are constantly subjected. Typical of these latter are warm drinking water, drafts, poor illumination, inadequate lockers, the number and placement of the time clocks; even such items as the fact that in the company cafeteria supervision may get larger helpings than do rank and file employees. The reason why these trivial and often picayunish complaints are of such major importance is the fact that *they are repeated every day*; they are inescapable. Furthermore, because they cannot "talk them out" with immediate supervision, and there are no channels of communication with top management, these dissatisfactions tend to accumulate and create anti-company attitudes. This is a serious causative factor in labor trouble.

The opinion poll also has a salutary effect upon employee morale in another direction: It provides the workers with an outlet for their dissatisfactions, thus serving as a cathartic agent. Having gotten their troubles off their chests, they feel much better. Likewise, the fact that management has evidenced sufficient interest in them as individuals to provide them with an opportunity thus to express their likes and dislikes tends also to create a bond of sympathy and goodwill between them and management.

The major resistance to the use of employee opinion polls arises not from employees or labor organizations, as might be expected, *but from management itself*. In one case in the author's experience a C.I.O. local itself initiated a request for an employee opinion poll in a labor dispute and agreed to abide by the findings. Management, on the other hand, flatly refused to permit such a poll to be conducted, even though the work was to be done by an independent, outside organization.

In view of the contributions which the opinion poll can make to employee morale it is, at first glance, curious why it encounters so much resistance by top management. Actually, however, the explanation is relatively simple. For all of their prestige and authority, members of management are frequently extremely insecure.

Many are surprisingly lacking in self-assurance, the feeling of being on top of their jobs; some consciously, others so beset by anxiety that they cannot face their weakness and must repress the entire conflict. In view of this, they dare not approve the use of such an instrument, the employee opinion poll, which might reveal *their own shortcomings*. In short, many lack the courage to face unpleasant and uncomplimentary revelations relative to the kind of a job they are doing. Unfortunately, such polls again and again reveal conditions which management prefers not to face. Furthermore, the very act of asking for employee opinions also commits management to take action to correct conditions found to be bad when it would be easier not to. Typical of the conditions an employee poll is apt to reveal are: 1. Poor operating methods; 2. Undesirable working conditions; 3. Weaknesses in supervision; 4. Inconsistencies and inequities in company policies; and 5. Hostilities toward top management itself. Inasmuch as many of these reflect directly on management's competence, it is obvious that many executives are not eager to have them brought to light.

Under such circumstances management usually feels it safest to let sleeping dogs lie. At the same time, however, management cannot admit or may not even be consciously aware of the true reasons for its reluctance to permit the conduct of an employee opinion poll. Hence its response to this threat to its security is to provide plausible rationaliza-

tions for its attitude. Where management's anxieties are so powerful that they cannot be faced at all, the excuses given for refusing to permit an employee opinion poll are sincerely believed by the executives themselves.

The most common reasons advanced by management for refusing to permit employee opinion polls to be conducted are:

1. The poll will "suggest" dissatisfactions, thus *creating* poor morale and ill-will toward the company.
2. The conduct of the poll will upset employees emotionally, distract them from their work, and result in much discussion of these matters both on and off the job, to the disadvantage of operating efficiency; there is both a direct and indirect loss of production time.
3. Many employees will refuse to answer the questions or will give silly or irrelevant responses.
4. The bringing of issues such as those covered by an opinion poll out into the open will give the unions ammunition to attack the company and even lead to open outbreak of labor troubles.

Actually, none of these criticisms is likely to be valid. Numerous polls have been conducted in both large and small organizations throughout the country. Where they have been properly administered, none of the conditions which management fears have developed. Specifically, the experience of companies which have used them properly has been as follows:

1. *The charge that polls create dissatisfaction:* This criticism has been proved false by two factors: First, it is almost always found that there are marked differences in the kind and degree of dissatisfaction found from department to department. If the poll itself were creating this dissatisfaction, there would be a much greater degree of uniformity. Second, a further more detailed investigation of the dissatisfactions voiced on the poll has revealed almost in every case that *they actually exist*. In short, the polls bring out only real dissatisfactions which already exist and do not cause employees to imagine new ones.

2. *Polls require too much time and distract employees:* Even the longest polls require not in excess of thirty minutes to explain, administer and collect. Rarely is there much discussion after the poll has been conducted. The fact that the employees have had an opportunity to get their dissatisfactions off their chests has the effect of reducing existing emotional tensions. They actually feel better. It is true that in some instances there is a certain degree of "kidding" of supervision after the poll has been taken, e.g., an employee will go to his supervisor, if he likes him, shake his hand, and say "Glad to have known you because, after manage-

ment has learned what we think of you, you won't be with us any longer." This actually is a sign of good morale. Rarely, if ever, are employees upset by a poll.

The executives who decry a loss of production time are obviously sticking their heads in the sand. They refuse to face the obvious fact that improved employee morale invariably means increased productivity. Their motives become transparent when one observes how quickly they buy costly plant equipment based on this same argument of future productivity.

3. *Employees will not answer the questions or will give irrelevant responses:* If the purpose of the poll has been sufficiently explained in advance, not to exceed one or two per cent of the employees will give "smart aleck" responses, and practically none will refuse to answer. Even where employees are strongly anti-management, the opportunity provided by the poll anonymously to tell management what they think of the company is so attractive that the individual can rarely resist the opportunity to unburden himself in full. Where irrelevant and "smart aleck" responses are written in, this in itself is an indication of serious hostilities toward, or distrust of, management.

4. *They will lead to outbreaks of labor troubles:* Opinion polls are ordinarily never discussed with the union in advance of their administration (otherwise they might be "framed"). This is sometimes resented, but at least in the author's experience, no reluctance is encountered on the part of union members to answer the questions, nor are there unfortunate consequences resulting from the administration of the polls. In one instance, the president of a local union announced that no member of the union would answer the questions on the poll. In spite of this, every member responded in full and even the president once he saw the questionnaire, found it impossible to resist the temptation to tell the management in strict confidence what he thought of it.

In another case, a group of organized railroad employees were given an opinion poll. Here each department had a "griever," and management feared that immediately each griever would call headquarters and the employees would be called on strike. Actually, not a single griever called the general chairman, and the immediate effect of the poll was a marked and consistent improvement in employee attitudes.

As already indicated, at least in one case, to the author's knowledge, a union has asked for a poll. In short, if the labor organization has confidence in the integrity and essential fairness of the company, there is no reason to anticipate trouble of any sort. The best proof of this is the fact that polls of this character have been conducted in numerous companies which have been organized by strong, aggressive unions, and

in no instance, in the author's experience, has trouble resulted from the taking of the poll.

Since the basic reason for management's resistance to opinion polls lies in its own anxieties, the only sure way to overcome it is to allay management's fears for its own security. Four methods have proved helpful in dealing with this problem:

1. *By referring management to executives of other companies which have conducted these polls successfully.* If executives in these latter companies explain that, at least in their experience, the polls not only do not cause trouble, but are sources of valuable information, this is extremely helpful in allaying management's fears. It is particularly advantageous to arrange contact with companies which have conducted not one, but a series of these polls, because this practice tends to give clear evidence that the polls are not only free from danger, but actually contribute information which is helpful.

2. *By making it clear to management that the findings will be handled confidentially.* In short, only those executives who are most likely to be injured by the findings will be shown the complete report. While it is essential that groups of employees be informed of those particular findings which relate to them, it is not necessary for them to be given the overall picture. Consequently, if conditions are unusually bad, no one but top management need know the extent and sources of employee dissatisfaction. In this manner, management is enabled to save face, while at the same time the findings may be employed constructively.

3. *Where the poll is recommended by an outside consultant, he can accept the responsibility for the effect of the poll upon the employees.* If he is willing to stake his reputation upon the outcome of the poll, management is sometimes willing to take the chance.

4. *Sometimes management will consent to the trial of a poll in one department on a pilot basis.* Later, reassured, it will go ahead with the entire organization.

In actual practice there is only one danger associated with the use of employee opinion polls. This is the failure of management to do its part in the correction of conditions revealed by the questionnaire. Frequently, a poll of this character brings out conditions which are either embarrassing to management or presents it with rather difficult problems. For example, the employees of a particular department may express strong hostilities toward their foreman. They may charge him with incompetence, with playing favorites, even with open dishonesty. On the other hand, this man may have been with the company for many years. In addition, the management may regard him quite highly because his costs are low and he is an excellent technician. Yet in spite of

these favorable factors, it becomes incumbent upon management either to give this supervisor special leadership training, or, in extreme cases, to transfer or replace him. This latter may confront management with a serious problem, either because of the man's length of service or because it has no one with whom to replace him. *In spite of this, it is essential that management if it is to hold the confidence and respect of its employees, must take action.* If the executives dodge the issue and take no action, hoping wishfully that the employees will forget their dissatisfactions or that the man will reform voluntarily, they will be doing themselves a great disservice. Not only will the initial cause of the dissatisfaction, i.e., the department head or foreman, remain, but a *new* ground for distrusting management will have been established. The employees will feel with justification that management has acted in poor faith. It has asked them to give their frank opinions with the implied promise either that conditions would be corrected or that at least management would give them an explanation of why action has not been taken. Where nothing is done, employees rightfully feel that they have been given the run-around and that their confidence has been abused. Under such circumstances subsequent employee opinion polls or other management projects will not be accepted readily by the workers.

On the other hand, if company executives will be honest with themselves and with the employees and make a genuine effort to eliminate the conditions revealed by the opinion poll, or if this is not possible, have a frank and open discussion with the men and women involved and give them an explanation of why changes cannot be made, the effects are rarely other than desirable. The employees have, often for the first time, tangible evidence that management is sincerely interested in their welfare, and is honestly attempting to improve conditions. Where this is done, not only are many of the basic causes of dissatisfaction eliminated and the employees given an opportunity to relieve their tensions by expressing them on the questionnaire, but an atmosphere of mutual confidence is established which is of primary importance in building and maintaining good morale.

Thus, where an employee opinion poll is properly conducted and full use is made of the findings, its effects cannot be otherwise than helpful, not only from the standpoint of the building and maintenance of employee morale, but in such by-products as improving levels of production and reducing absenteeism and turnover, since the latter conditions arise in large part from unrelieved employee dissatisfactions.

Received March 20, 1946.

Signed Versus Unsigned Personal Questionnaires

Robert P. Fischer

University of Illinois

Despite considerable literature on questionnaires, personality inventories, attitude scales, etc., there is a paucity of empirical data on the influence of signatures upon the results obtained with these devices. Many writers have recommended keeping questionnaires, etc., anonymous, but presemably have done so on the basis of personal conviction rather than on a basis of any factual observations. Usually it has been asserted that where the information sought is of a personal nature, or, as in the case of an attitude study, where the respondent's views are likely to be in disagreement with those held by the examiner, anonymity is essential in order to obtain honest information. Such claims are, however, for the most part unsubstantiated.

The present writer is aware of only two studies in which the influence of signatures upon questionnaire data has been investigated. Olson studied the influence of waiver of signature on personal reports. Using the Woodworth-Mathews Personal Data Sheet, which is designed to measure emotional stability, Olson examined 100 upperclass women who were preparing to be teachers. One group of 60 women was given the data sheets with instructions not to sign them. Immediately after finishing the task they were instructed to again fill out the data sheet but this time to sign their names. These conditions were reversed for another group of 40 students. On the basis of the results he obtained, he concluded that, "There is thus a high probability that more symptoms will be reported in an initial application of the instrument when names are omitted." He further concluded that, "The initial experience, however, appears to establish a set or memory factor which prevents large changes on the second application to the same group."

Corey² in a study of the influence of signatures on attitude questionnaires, however, did not get the same results. He used a questionnaire designed to measure attitude toward cheating on examinations. By use of concealed pin pricks he was able to identify the anonymous

¹ W. C. Olson. The waiver of signatures in personal data reports. *J. appl. Psychol.*, 1936, 20, 442-450.

² S. M. Corey. Signed vs. unsigned attitude questionnaires. *J. educ. Psychol.*, 1937, 28, 144-148.

papers of the 150 college students used in his study. He had the students first respond to the questionnaire but not sign their names. After the papers were collected he had the group respond in the same way except that he had them sign their papers. He discovered no statistically significant differences between the mean scores obtained under the two conditions of administration. The reliability of the questionnaire was equally high under both conditions being .93 for the unsigned questionnaires and .90 for the signed ones. The coefficient of correlation between the scores made under the two conditions was .85. From these results he concluded that, "—the concern of investigators over the invalidating effects of a signature may have been exaggerated."

Several factors may have been operating to make the results of these two studies dissimilar. In the first place, there may have been some differences in the subjects used. There may have been differences in the set of the two groups. Olson's subjects, it will be recalled were upper-class women preparing to be teachers. It is possible they may have tried to anticipate the nature of the experiment and thus reacted differently from Corey's subjects. Then too, there was a difference in the tests used. Whatever may have caused the variation in results in the two studies reported above, the problem of the influence of signatures on questionnaire type data still needs some investigation. The present study was carried out to throw further light on the problem. The College Form of Mooney's Problem Check List was administered under two sets of conditions, first with signatures and then without, to obtain the data for this study.

The College Form of Mooney's Problem Check List ³ is designed to aid students in the expression of their personal problems. It consists of 330 items and 5 questions. Each item is intended to suggest a possible personal problem, while the 5 questions are included for summary purposes. The 330 items are classified under eleven general headings which are shown in Table 1. The student is instructed to read through the list of items, underline those which are of concern to him and then to go back over the list of underlined items and circle those which are of most concern to him.

There is a place provided on the check list for the student's name but whether he signs the list depends upon the use that is to be made of it. When some counseling follow-up is intended the student is instructed to sign the list. If the data are to be used in group form, for example in surveying the problems of a specific class, signing is presumably regarded as inadvisable. It was in order to determine the influence of signing the

³ R. L. Mooney. *Problem check list*. Columbus, Ohio: Bureau of Educational Research, Ohio State University; 1941.

check lists on the number of items underlined and circled that the following study was carried out.

Conditions of Present Study

The Mooney check list was given to the students in two of the writer's classes in psychology at the University of Illinois. They were administered at the beginning of the eighth week of the fall semester, 1944-45. The responses of 56 sophomore women in a class in general psychology and 46 upperclass women (mostly juniors and seniors) in a class in industrial psychology were used in this research. The students in the class in general psychology were well acquainted with each other having been in several other classes together. While this was not true of the women in the class in industrial psychology, every effort was made to see that these students became acquainted with each other during class. The degree of rapport with both classes was high. The writer had had conferences with most of the students and had tried in every way to establish friendly relations with them. Compared with previous experience with other classes the cooperativeness of these two groups of students was outstanding.

In both classes the students were on topics at the time of participating in this research which made the use of the check list appear quite in order. Both classes were studying personality and its measurement in the general context of personal adjustment.

The check lists were given to the two groups of students on one Monday with the following instructions: "You have been studying personality and its measurement for over a week now. One of the best ways to really learn about the measurement of personal adjustment is to take some of the tests we have been discussing yourselves. Therefore, today I want to give you the College Form of Mooney's Problem Check List (this check list had not been discussed previously). Now this is not a test in the usual psychological sense but consists of a number of things which might be problems to you. I want you to read the instructions on the check list and then proceed to fill it out as honestly and frankly as you can. Your papers will be kept strictly confidential and later I will discuss them with you individually. You should get at least two benefits from doing this project. First, you should learn something first hand of the nature of problem check lists and second, you should gain some insight into the things that are disturbing you. Please be as frank and honest as you can or the results won't be of any value to you. Remember, no one else besides me will see your papers." They were then told to proceed.

On the following Monday the check list was again given to the same students.⁴ This time it was given with the following instructions: "You will remember that a week ago you each filled out a problem check list. I have gone over them and am ready to discuss them with you. In scoring them

⁴The students who were not present at the first administration were dismissed before class began. The interim of one week should avoid the memory factor pointed out by Olson and probably also present in Corey's results.

I noted some interesting trends and would like to make some group summaries. Of course I will not turn your signed check lists over to an assistant to work on but I do want them summarized. I wonder if you will fill one out today but this time do not sign it. This way my assistant can work on them and not know whose they are. Please be as frank and honest as you can and do not sign the check lists or mark them in any other way that may identify them." The students were then told to proceed. The writer then left the room in charge of an assistant to allay any suspicion that the papers might be identified.

During the interim between testings no mention was made of the results of the first administration of the check list and no other check list or personality tests were given. In fact, the classes were, for all intents and purposes, through with the topic of personality.

Results

A preliminary analysis of the results made by the two classes, both when the check lists were signed and when they were not, showed that they did not differ appreciably. Significance tests were applied to the small differences obtained between the two classes but in no case were the differences statistically significant. Accordingly the papers of the students in the class in general psychology were combined with those of the students in the class in industrial psychology into a single group of papers. There were 102 such papers on the first testing (with signature) and 94 on the second testing (without signature). This difference of 8 was due to an absence of 8 students at the time of the second testing who were present at the time of the first testing. There is no reason to suspect that these 8 cases would have altered the results any. Unfortunately, since attendance was not taken (to further allay suspicion) it was not possible to identify the 8 missing papers and thus the papers of these absentees were not removed from those obtained at the first testing.

The mean number of problems underlined when the check lists were signed was 34.37 while the corresponding mean when the check lists were unsigned was 36.00. This difference of 1.63 had a critical ratio of 0.54. Obviously this was not a statistically significant difference. The mean number of problems circled (serious problems) on the signed papers was 8.11 and 11.32 on the unsigned papers. This difference of 3.21 had a critical ratio of 2.38. While a critical ratio of this magnitude is not the conventional 3.00 often demanded for statistical significance, it is great enough to indicate a difference in directionality of the above two means which is probably due to something other than chance. It shows that there tended to be significantly more serious problems on the unsigned check lists than on the signed ones.

Table 1 shows the mean number of problems underlined (total problems) in each of the eleven areas on both the signed and unsigned check

Table 1
Problems Underlined by Group With and Without Signatures

Problem Areas	Mean Names	Mean No Names	$M_1 - M_2$	Critical Ratio
Health and Physical Development	2.863	2.734	-.129	-.374
Finances, Living Conditions, and Employment . .	1.294	1.479	.185	.703
Social and Recreational Activity	3.324	3.969	.645	1.352
Social-Psychological Relations	3.431	3.192	-.239	-.508
Personal-Psychological Relations	4.794	4.767	-.027	-.048
Courtship, Sex, and Marriage	3.147	3.490	.343	.940
Home and Family	2.186	1.873	-.313	-.801
Morals and Religion	2.539	2.745	.206	.544
Adjustment to College Work	3.804	4.182	.378	.696
The Future—Vocational and Educational	5.029	4.980	-.049	-.092
Curriculum and Teaching Procedure	1.961	2.586	.625	1.586
Total ¹	34.372	35.995	1.623	.543

¹ Results obtained from the original data, not from the figures shown above.

lists. It further shows the differences in the above means along with the critical ratios of these differences. Table 2 shows comparable data for the circled problems.

From Table 1 it may be seen that in none of the eleven areas were there any significant differences in the mean number of problems underlined by the group with as compared without signatures. In general it may be assumed that whether or not the group signed the check list was of no importance in determining the average number of problems underlined in the eleven areas. It will be noted that the greatest differences were in the areas headed: "Curriculum and Teaching Procedures," and "Social and Recreational Activities," but these differences were not statistically significant.

From Table 2 it may be seen that there were fairly significant differences in the number of problems circled (serious problems) in three of the areas by the group with as compared with the group without signature. There was a tendency for a higher average number of serious problems in the areas headed, "Curriculum and Teaching Procedures," "Finances, Living Conditions, and Employment," and "Social and Recreational Activities," when the students did not sign their check lists than when they signed them.

In general, therefore, it may be assumed that the students in this study tended to circle more problems when their names were withheld from the check lists than when their names were used, but that there was no significant difference in the number of problems underlined under the

Table 2
Problems Circled by Group With and Without Signatures

Problem Areas	Mean Names	Mean No Names	M_1-M_2	Critical Ratio
Health and Physical Development569	.713	.144	1.021
Finances, Living Conditions, and Employment . .	.225	.479	.254	2.153
Social and Recreational Activity529	.883	.354	1.924
Social-Psychological Relations588	.766	.178	1.066
Personal-Psychological Relations	1.206	1.457	.251	.893
Courtship, Sex, and Marriage990	1.425	.435	1.851
Home and Family549	.596	.047	.220
Morals and Religion549	.840	.291	1.516
Adjustment to College Work	1.108	1.691	.583	1.685
The Future—Vocational and Educational	1.549	1.840	.291	.977
Curriculum and Teaching Procedure245	.628	.383	2.424
Total ¹	8.108	11.319	3.211	2.384

¹ Results obtained from the original data, not from the figures shown above.

two conditions. In view of the relatively small sample used and the special conditions of rapport, it is not possible to assume that the differences observed were necessarily due to the withholding of signatures nor would it be reasonable to generalize these results to other populations. Nonetheless, the fact that the use of signature on personal questionnaires may influence the honesty and frankness of students, as indicated by the number of personal problems checked, seems a distinct possibility. Further evidence for this possibility lies in the similarity of the results of this study with those of Olson noted earlier.

Summary

The College Form of Mooney's Problem Check List was given to 102 upperclass women students in psychology first with and then without signatures being used. The interim between testings was one week. The results indicated that the mean number of problems underlined (total problems, presumably not serious) did not vary significantly under the two conditions of administration but that the mean number of problems circled (serious problems) tended to be significantly greater when signatures were withheld. In view of similar results reported by Olson it would appear that the use of signatures on personal questionnaires (particularly in the case of highly personal items or serious problems) might have a relative inhibitory effect on the honesty and frankness of the people responding to them.

Received July 11, 1945.

Age of College Graduation and Success in Adult Life *

S. L. Pressey

Ohio State University

In discussions of educational acceleration it is often argued that although young students may make admirable college records, early graduation starts them into their adult careers while yet so immature, or they are so often a "bright boy" type of personality, that early promise is not fulfilled. Instead, the graduate at average or older age is supposed to have a maturity, and a greater variety of experience as in work or travel, which results in a more substantial adult career. Now there is the special problem as to whether the veteran who returns to school, and finally gets started in his civilian career older than usually occurred before the war, will gain or lose because of his greater age; if the latter, accelerated programs for veterans are suggested, and as a matter of fact are desired by many veterans. The data here reported bear on these various related matters and seem especially timely now.

Material and Methods

The issue was straight-forward and specific: What is the success in life-career of students who graduate from college young, at an average age, or older? For study of the problem, alumni records of as great completeness as possible seemed needed. In most institutions such records are notoriously inadequate and especially so with reference to the careers of the less successful graduates. However, Amherst College has notably complete published alumni records from the first graduates to 1938, including for almost all former students their vocational careers, status in community or profession as shown by honors or other recognition, memberships in social or community or professional groups, and family. Sources other than the individuals concerned seem often to have been used, and exceptional adequacy and accuracy obtained. The volume thus seems a mine of information for studies regarding certain problems in higher education.

[The purpose of the investigation here reported was to appraise total adult careers. Only classes the careers of whose members had been largely completed at the time when the data in this volume were gathered,

* This is the 25th in a series of reports regarding research in the Bureau of Educational Research of the Ohio State University, regarding educational acceleration.

could therefore be used. But cases were desired as close to the present as the above requirement permitted. Various considerations led to use of the classes of 1880 through 1900. Clearly an individual who died young did not have a chance to show what he could do. Only graduates in these classes who had lived to be 50 years or over were accordingly included and only those born in this country. The question is as to whether or not age of college graduation shows any relationship to the adult careers of these individuals.

Results

Table 1 shows for graduates at each age the per cent not marrying, the average age of first marriage for those who did, and the average number of children for both groups together. Youngest graduates are slightly more likely to marry and do so youngest; older graduates marry later but hardly less frequently. Average number of children runs practically the same throughout.

Table 1

Number Marrying, Age of Marriage, and Number of Children (entire group) in Relation to Age of College Graduation, 1411 American-born Graduates of Amherst in the Classes of 1880 through 1900 who lived to be 50 or over

Age of Graduation	19	20	21	22	23	24	25	26	27 Up
Number of Graduates	24	114	344	416	228	104	82	37	62
% Not Marrying	8	6	10	10	8	21	14	11	10
Av. Age of Marriage	26.8	29.9	29.8	30.6	30.5	31.1	31.5	31.0	34.1
Av. No. of Children	2.0	1.6	1.8	1.8	1.8	1.8	1.9	2.0	2.0

Average age of marriage of Harvard graduates of the classes of 1891-1900 was 31.0 and 25% of the marriages were childless (8). For the Amherst group here dealt with average age of marriage was 30.5 and 24% were without children. A more recent study (1) shows 9.2% of male college graduates over 40 in this country never to have married. In short, the Amherst data in total appears to be reasonably typical, (in these and yet other ways which need not here be gone into). It seems a fair inference that the striking findings next to be presented are of some general significance.

The material of major importance must now be considered, the findings as to adult career. Careers were appraised on a scale of seven. Individuals who were internationally known were rated "7"; "6" indicated national prominence and "5" prominence locally; "4" was for average success for a college graduate and "3" for only a mediocre career;

"2" indicated relatively unskilled work, "1" a failure, (not self-supporting), and "0" a criminal or shady record. The first two categories were found relatively easy to assign, such criteria being used as inclusion in "Who's Who." Local prominence was considered to be indicated by membership in local or regional organizations or similar evidence. At the other extreme, no case was found with a criminal or shady record or practically throughout not self-supporting; such cases either did not occur or were mercifully not so reported. However, an appreciable number were classified as "2"; thus one man after rather obviously failing in newspaper work had spent most of his life in relatively unskilled factory work, while another had gradually dropped back and spent the last 25 years as a letter-carrier.

Procedure in rating was simple. The statement about each individual was read and rated by at least two assistants, neither for the most part knowing the basic purpose of this study or paying attention to age

Table 2

Adult Careers of 924 American-born Amherst Graduates Who Lived to be 50 or Over:
All Those Graduating Under 21 and Over 25 Between 1880 and 1900
and the Entire Graduating Classes in 13 of These Years

Age of Graduation	19	20	21	22	23	24	25	26	Over 26
Number of Graduates	24	114	216	235	132	59	47	37	60
After College Success									
% Nationally known	29	22	15	12	10	3	2	3	
% Failures	4	6	6	5	2	3	6	11	15

of college graduation. If the same rating was assigned by both appraisers, it stood; if the ratings differed by but one, the lower of the two ratings was used as probably nearer right in view of the tendency of such statements to be over-favorable. If the raters differed more, final decision was made by a third person in close touch with the study. Such ratings were made for all cases graduating under 21 or over 25, in the 21 graduating classes from 1880 through 1900. To save time, no ratings were made on those graduating at the most common ages of 21-24 in the classes of '83, '84, '88, '89, '93, '94, '97, and '99. The cases at these ages from the other 13 classes seemed sufficient. On the whole, it is believed that these ratings appraised the life careers of these individuals with reasonable accuracy, especially at the extremes shown in Table 2.

The relationship of age of graduation to later success is surely striking. A quarter of the small group graduating at 19 was nationally known. And the per cents thus known drop steadily until no such cases

are found for those graduating over age 26. Moreover, failures are not more common among the youngest graduates; they are not more often maladjusted or unstable nor do they often, like Williams James Sidis, eventually become dismal failures. The number of failures is slightly less at 23 and 24, one might hypothecate that these were the regular ages when the conventional best adjusted students came through. But the differences are so slight as to be of doubtful significance, and it would seem best to infer simply that among graduates from 19 through 25 the per cent of failures is about the same. After that, failures become more common.

May the above findings result simply from the selective processes going on throughout the educational system, bringing it about that the brightest students get into and through college youngest, whereas mediocre or dull individuals enter college a year or two late because of poor work in elementary or secondary school, or in college take longer than the usual four years? The last possibility seems for the most part not a factor here, because the Amherst alumni records seem to list a man as in the class with which he entered unless he specifically requests otherwise. Thus a man who entered in 1880, but took five years because of poor scholarship, would be listed with the class of '84 and not '85. But late entrance due to poor ability might well operate.

Forty years and more ago, colleges did not give intelligence tests. But a little help can be got, for judging possible relationships of ability

Table 3

Age of Graduation and General Ability as Tested at Entrance, 1096 Graduates from 5 Undergraduate Colleges of Ohio State University, School Year 1941-42

Percentile	Under 21	21	22	23	24	25 Up	Total
90-100	24	111	97	29	20	30	311
80-89	5	56	58	29	12	20	180
70-79	5	48	51	21	11	21	157
60-69	1	39	48	23	11	13	135
50-59	2	29	26	22	10	7	96
40-49	1	18	31	11	9	9	79
30-39	2	21	21	17	4	8	73
20-29	1	9	13	7	2	7	39
10-19	1	5	7	1	3		17
0-09		2	3	1	2	1	9
Md.	91	80	76	69	70	76	
No.	42	338	355	161	84	116	1096
% O. S. U.	4	31	32	14	8	11	Md. 22.5
% Amherst	11	27	26	15	8	13	Md. 22.4

to age of graduation, by seeing what they are now, and considering whether the situation at Amherst might not have been similar. Table 3 attempts this comparison. It shows general ability at entrance, as measured by the Ohio College Association Test of General Academic ability, for 1096 graduates in the five under-graduate colleges of Ohio State University in the school year 1941-42.

As the medians show, those entering at 20 or younger made somewhat higher average scores on the scholastic ability tests at Ohio State University than those graduating 25 or over. The differences are not very great, 15 percentiles between youngest and oldest groups. The percentile form of statement, however doubtless conceals the marked superiority of some of the youngest students. But many of the older students also were of very good ability; over a quarter of those graduating 25 and older tested in the upper tenth. The two bottom rows show the distribution of graduation ages at Amherst and Ohio State University to be quite similar except for a somewhat larger proportion of very young and also older graduates at Amherst. The inference then is that the more frequent successes among the young graduates and more frequent failures among the older are not due entirely to differences in ability between older and younger groups.

Interpretation and Application

How then are the findings regarding relation of success to age of college graduation to be construed? There undoubtedly are relations to ability, as discussed in the preceding paragraphs. But that factor seems not enough. Presumably, somewhat related socio-economic influences play a part. The youngest graduates are likely to be those from homes of means and of advantages both facilitating education and furthering the beginning of life career. The writer, however, believes that there are in this situation two other factors now largely neglected, and the second of great possible importance.

In the first place, the older (not the younger) students are often in various ways maladjusted. Practically every investigation bearing on the matter gives support to this conclusion. Typical are certain of the writer's findings based on the records of 5,977 freshman entering the five undergraduate colleges of Ohio State University, and 2,055 graduates. Only 19 per cent of those entering over 21 graduated in the regular four academic years as compared with 33 per cent of those entering at 16; only 18 per cent of graduates over 24 had academic records averaging "B" or better as compared with 33 per cent of graduates under 21; 34 per cent of graduates over 24 did not participate in any way in extra-curricular activities as compared with only 10 per cent of those graduat-

ing below 21, and almost three times as many of the younger as compared with the older group had held office. The figures regarding activities seem especially important as indicative of difficulty in adjusting to campus life. All this may be due again primarily to less favorable socioeconomic status of older students, or to interruptions in schooling. But the writer ventures the hypothesis that underlying developmental factors, which have unfortunately been almost completely neglected by both psychologists and educators, may be yet more important. Surely physiological changes around the age of twenty, when the organism stops growing, are in total more profound than the special and comparatively more incidental changes of puberty, when it is well recognized that repercussions in the personality are great. Much scattered evidence suggests that changes in personality around twenty are also marked. The individual becomes more serious and purposeful, desires to marry, to establish himself as independent of his family and of tutelage from the previous generation. The older student feels belated and out of place with youngsters still in later adolescence, and uncomfortable that he is not yet into recognized independent adulthood intellectually, socially, or economically. All this might contribute to the maladjustment of the older student while in college, and also handicap him as he moved into his career after graduation.¹ Surely such incongruities between developmental level and student status are likely to be more acute with returning veterans, who will be older in years and much more in experience than the average pre-war student.

The second factor may well be much more important. It too can only briefly be touched upon here but is systematically discussed in reference 9. Recent research has emphasized that the prime of life in health and vigor, in intellectual creativeness, in energy and enthusiasm, comes early in adult life. Outstanding discoveries in science, finest books in literature, greatest inventions, all such accomplishments tend to be made by relatively young men (2, 3, 5, 6, 9, 10). In short, it may be far more important than is ordinarily realized, that a man be started in his life career early in his prime, if his maximal potentialities are to be realized. And even a few years delay may well be serious.

At least such are possibilities by way of explanation of the above data. And both the data and considerations such as have been mentioned above seem of especial importance now. The colleges are now dealing with large numbers of much older men returning from the services or war work. Furthermore, a year of military training may be required, again probably delaying college graduation and the beginning of adult careers. If the

¹ The issues of necessity only briefly touched upon in this and the following paragraph are more adequately dealt with in reference 9 and given larger perspectives in 10.

above findings indicate a major relationship between age of completing full-time education and adult success, issues are presented of the greatest importance, both for education and for national policy as regards utilization of the human resources of the ablest group of young people in the country. Acceleration, especially of veterans but also of able students generally, seems called for. Finally, peace-time conscription appears to entail problems ordinarily neglected.

Summary

1. The paper reports an effort to determine relations between age of college graduation and success in adult life. The unusually complete alumni records of Amherst College for graduates in the years 1880-1900 inclusive, who were born in this country and who had lived to be at least 50, were the data for the study.

2. For these classes, it was found that the per cent marrying and number of children did not vary significantly with age of college graduation. Age of marriage, however, increased with increased graduating age.

3. Most significant was the relationship of graduating age to vocational success. Success was judged by ratings by two raters working independently with arbitration by a third in case of disagreement; the ratings seemed of considerable reliability. A steady decrease in per cent highly successful, nationally or internationally known, was found from youngest graduates to oldest. Those graduating at an older age were more likely to have been failures.

4. Two special hypotheses are offered in explanation: (a) Older students were maladjusted to college work and college life, with consequent handicap in adulthood. (b) Late graduation too much reduced the number of most vigorous years in the prime of life which might have gone into most energetic initiation of life career. It is believed that such findings argue for a judicious acceleration of educational programs, especially for veterans, and argue against peacetime conscription.

Received April 21, 1945.

References

1. Babcock, F. L. *The United States college graduate*. New York: Macmillan, 1941. Pp. 112.
2. Collins, S. D. A general view of the causes of illness and death at specific ages. *U. S. Public Health Report*, 1935, 50, 237-255.
3. Collins, S. D. Cases and days of illness among males and females with special reference to confinement to bed. *U. S. Public Health Report*, 1940, 55, 47-93.
4. Keys, N. The under-age student in high school and college. *Univ. Calif. Publ. Educ.*, 1938, 7, 145-272.

5. Lehman, H. C. The creative years in science and literature. *Sci. Mo.*, 1936, **43**, 151-162.
6. Lehman, H. C. Man's most creative years. *Sci. Mo.*, 1944, **59**, 384-392.
7. Newlin, William (Editor). Biographical record of the graduates and non-graduates. Amherst College, 1939, p. 953.
8. Phillips, J. C. Further studies of the Harvard birthrate. *Harvard Grad. Mag.*, 1925-26, **34**, 385-394.
9. Pressey, S. L. A neglected crucial psycho-educational problem. *J. Psychol.*, 1944, **18**, 217-234.
10. Pressey, S. L., Janney, J. E., and Kuhlen, R. *Life: A psychological survey*. New York: Harper, 1939. Pp. 654.

Predicting Success in a School of Nursing *

A. Q. Sartain

Southern Methodist University

It is important to a school of nursing to be able to predict, with as great accuracy as possible, the success that an applicant for admission will have in the school, and to be able to choose as nearly as possible only those likely to succeed. Not only is elimination of those likely to fail a service to the weak applicants, but it is usually decidedly to the advantage of the school of nursing, since, as Potts¹ has pointed out, students are usually liabilities rather than assets to a hospital until some months have passed.

Statement of the Problem

The purpose of this study was to determine the extent to which success in nursing school could be determined from the high school averages of students and also from their scores on a battery of tests administered by the Nurse Testing Division of the Psychological Corporation (the Revised Alpha Examination, Form 8; the Columbia Vocabulary Test; the MacQuarrie Test for Mechanical Ability; the Bernreuter Personality Inventory; and the Potts-Bennett Tests for Nursing Aptitude—referred to in this study as though the two sections comprised a single test).

History of the Problem

A number of studies have concerned themselves with the use of psychological tests in the selection of nursing school candidates, and at least two agencies, the Psychological Corporation and the National League of Nursing Education, administer their test batteries to many applicants each year. The battery used by the Corporation has just been described,

■ The writer wishes to express his appreciation to Miss Merle Mayo, R.N., of the Parkland Hospital School of Nursing, who furnished the grade average and the high school average for each girl and whose cooperation in obtaining the test scores made this study possible; to Mrs. Margaret Scruggs-Carruth, who did a great deal of the statistical and other work; and to Miss Edith M. Potts, R.N., of the Nurse Testing Division of the Psychological Corporation, who encouraged and cooperated with the study.

¹ Potts, Edith Margaret. Use of tests in selecting student nurses advantageous to hospital and student. *Hospital Management*, 1941, 52, 39-42.

and the League makes extensive use of tests of the Cooperative Test Service and also employs the A. C. E. Psychological Examination.

One of the most extensive investigations in this field was the work of Williamson, Stover, and Fiss.² Although handicapped by a lack of records on students who did not complete a full year of work, and by varying standards of grading, they found that the Moss Nursing Aptitude Test, the Cooperative English Test, the Cooperative General Science Test, and the Cooperative Vocabulary Test correlated best with nursing school success, with multiple regression coefficients reaching .54 in some cases. A similar investigation by MacPhail and Bernard³ yielded coefficients of correlation between intelligence test scores and preliminary grades of from .42 to .60, though differences between those graduating and those failing to do so were not significant in two of the four schools studied. Douglass and Merrill⁴ secured a multiple regression coefficient of .77 when success in a school of nursing was predicted from scores on the Moss Nursing Aptitude Test, the Cooperative General Science Test (Part I), the Douglass-Gordon Fraction Test, and the high school percentile rank. The Moss Nursing Aptitude Test and the high school percentile rank yielded a coefficient of .75.

Rainier, Rehfeld, and Madigan⁵ obtained correlations of more than .40 between nursing school grades and the Iowa Reading Comprehension Test, while Garrison⁶ obtained correlations of .48 between academic grades and the Otis Self-Administering Test of Mental Ability and .59 between nursing arts grades and the Otis. Bennett and Gordon⁷ made a careful analysis of scores on the Bernreuter Personality Inventory and concluded that the test was of little or no value in predicting success in nursing school. Finally, Scruggs-Carruth⁸ made a preliminary study of the subjects employed in the present group.

² Williamson, E. G., Stover, R. D., and Fiss, C. B. The selection of student nurses. *J. appl. Psychol.*, 1938, 22, 119-131.

³ MacPhail, A. H., and Bernard, W. Ten years of intelligence testing. *Educ. & Psychol. Meas.*, 1943, 3, 157-165.

⁴ Douglass, H. R., and Merrill, R. A. Prediction of success in the school of nursing. *Univ. Minn. Stud. in the Prediction of Scholastic Achievement*, 1942, 2, 17-31.

⁵ Rainier, R. N., Rehfeld, F. W., and Madigan, M. E. The use of tests in guiding student nurses. *Amer. J. Nursing*, 1942, 42, 674-682.

⁶ Garrison, K. C. The use of psychological tests in the selection of student nurses. *J. appl. Psychol.*, 1939, 23, 461-472.

⁷ Bennett, Geo. K., and Gordon, H. Phoebe. Personality test scores and success in the field of nursing. *J. appl. Psychol.*, 1944, 28, 267-278.

⁸ Scruggs-Carruth, Margaret. The predictive value of nursing school tests. Unpublished Thesis: Southern Methodist University, Dallas, Texas. 1944.

Subjects and Conditions of the Study

Eighty-one girls from the Parkland Hospital School of Nursing in Dallas, Texas, comprised the experimental group. These girls took the Psychological Corporation tests in 1942, and were admitted to the School of Nursing regardless of their scores on any of the tests. The criterion of success was the average grade earned by the student in all courses by the end of six months of training, or at the time the girl left the School of Nursing, if she was not in school six months later. Sixty-nine of the girls were still in school at the end of six months, and twelve had left, in almost every case because of failing grades. The faculty members who examined the average grades thought them to constitute in the main accurate measures of actual success in the School.

Results of the Study

The only scores available on the Bernreuter Personality Inventory were percentile scores on Emotional Stability, Dominance, Extraversion, and Self-Sufficiency. Table 1 gives the results of correlating the scores on these traits with the grade average. Since these coefficients are

Table 1

Correlation Between Grades and Percentile Scores on the Bernreuter Personality Inventory

Trait	<i>r</i>
Self-Sufficiency29
Dominance26
Emotional Stability19
Extraversion17

based on percentile ranks and since they are low, as would be expected from other studies, they are not used further in this study. It is evident that they are not high enough to be of practical value, and there is little reason to believe that the use of the actual scores instead of percentile ranks would materially alter this situation.

One of the difficulties encountered was the determination of comparable high school averages for each girl,⁹ since grades might be in terms of letters or numbers or even descriptive adjectives, inasmuch as the girls came from widely scattered high schools. The method adopted for this study was to give an "A" (or other highest category) a value of 95 (with an "A plus" valued at 98 and an "A minus" valued at 92), a "B" a value of 85, and so on, and thus to convert all grades to a numerical basis. Needless to say, this introduced errors into the calculations, but obviously

⁹ No high school record was available for four girls. Consequently, correlations involving high school average were based on 77 cases only.

these differences in grading systems constitute one of the serious limitations of the high school average as a means of prediction. This can be overcome, however, when, as in Minnesota, Ohio, and other states, the high schools can be persuaded to report high school averages in terms of relative rank.

Table 2 gives the product-moment coefficients of correlation found by intercorrelating the variables. It should be noted that the test scores are total scores, no account being taken of the fact that some of the tests yield scores on two or more subtests.

Table 2

Coefficients of Correlation Between Variables, and Mean and Standard Deviation of Each Variable

Variable	Potts-Bennett	Army Alpha	Col. Voc.	H.S. Ave.	Mac-Quarrie	S. of N. Ave.	Mean	S.D.
Potts-Bennett	—	.776	.789	.429	.486	.677	165.1	42.40
Army Alpha	—	—	.776	.387	.393	.559	122.6	22.89
Col. Vocab.	—	—	—	.392	.365	.517	73.3	11.73
H.S. Average	—	—	—	—	.199	.460	84.3	5.52
MacQuarrie	—	—	—	—	—	.356	171.1	29.70
Sch. of Nurs. Av.	—	—	—	—	—	—	80.2	8.22

Table 2 reveals that the correlation between the criterion of success (school of nursing average) and scores on the Potts-Bennett test is quite high. The Alpha and Columbia Vocabulary tests also correlate with the criterion fairly highly, while the high school average and particularly the MacQuarrie do less well. Clearly, scores on the Potts-Bennett alone aid considerably in predicting success (improvement over chance 26.4%, S.D._{est} 6.05). And of course, as Tiffin¹⁰ suggests, if one can maintain a low selection ratio, the test has even more utility than these figures indicate.

An additional question concerns the possibility of improving prediction still further by the best combination of test scores and high school average. The multiple regression coefficient obtained from the values in Table 2 was .707 (improvement over chance 29.3%, S.D._{est} 5.81), and the formula for the best prediction of the criterion was (in terms of Beta coefficients):

$$\text{Sch. of Nursing Av.} = .584 \text{ P-B} + .103 \text{ Alpha} - .118 \text{ Col. Voc.} + .209 \text{ H.S. Av.} + .033 \text{ MacQ}$$

¹⁰ Tiffin, Joseph. *Industrial psychology*. New York: Prentice-Hall, Inc., 1942, p. 33.

Expressed in terms of raw scores the formula was

$$\text{Sch. of Nursing Av.} = .113 \text{ P-B} + .037 \text{ Alpha} - .083 \text{ Col. Voc.} + .311 \text{ H.S.Av.} + .009 \text{ MacQ} + 35.18.$$

Another multiple regression coefficient that was determined involved the prediction of the school of nursing average from the tests alone (without high school average). Here the coefficient was negligibly higher than for the Potts-Bennett alone, .680. And when the criterion was predicted from the Potts-Bennett and the high school average, the regression coefficient was .702, and the formula for prediction was (in terms of Beta coefficients):

$$\text{Sch. of Nursing Av.} = .588 \text{ P-B} + .208 \text{ H.S.Av.}$$

In terms of raw scores the formula was

$$\text{Sch. of Nursing Av.} = .114 \text{ P-B} + .309 \text{ H.S.Av.} + 35.28.$$

It is evident from these facts that the Potts-Bennett Tests alone do a creditable job in predicting success in this particular school of nursing,

Table 3

Coefficients of Correlation Between School of Nursing Average and Subtests of the Potts-Bennett and MacQuarrie Tests

Test	<i>r</i>
<i>Potts-Bennett</i>	
Science Information691
Paragraph Comprehension630
Specialized Vocabulary581
Speed of Reading352
Arithmetic Process467
General Information443
Arithmetic Reasoning472
<i>MacQuarrie</i>	
Speed and Coordination263
Mechanical Insight356

and indeed that neither the other test scores nor the high school average (nor both together) make any substantial improvement in this prediction.

Although they did not figure in the multiple regression coefficients referred to above, correlations were obtained between the school of nursing average and the subtests of the Potts-Bennett and the MacQuarrie. Table 3 gives these coefficients. From this table it appears that the subtests are no better than the complete tests in their predictive value, and there is little reason here to believe that a multiple regression

coefficient here would alter this conclusion. And of course, the lowered reliability likely to result from the use of any shorter test alone makes this course of action unwise at least until additional work is done on this point.

Summary and Conclusions

There were obtained, for approximately 80 girls entering a school of nursing, the high school average and scores on the five tests generally used by the Nurse Testing Division of the Psychological Corporation. These tests were the Revised Alpha Examination, Form 8; the Bernreuter Personality Inventory; the Columbia Vocabulary Test; the MacQuarrie Test for Mechanical Ability; and the Potts-Bennett Tests. The average grade earned in the school of nursing during the first six months (or up to the time of withdrawal if the girl did not complete six months of training) was used as the criterion of success, and these grades were correlated with the test scores and the high school average, and intercorrelations were determined for scores on the various tests (except Bernreuter) and the high school average. Correlations were also worked out for the subtests of the Potts-Bennett and the MacQuarrie. On the basis of the study the following conclusions seem to be justified:

1. The Potts-Bennett Tests were fairly effective in predicting success in the school of nursing, the coefficient of correlation being .677.
2. Addition of the other tests (not including the Bernreuter) to the Potts-Bennett improved the predictive value by a negligible amount ($R = .680$).
3. Addition to the above of the high school average yielded some increase ($R = .707$) but probably not enough to justify the time and energy involved. The high school average used, however, was not a measure of *relative* scholastic ability.
4. The Potts-Bennett Tests and the high school average yielded a multiple regression coefficient of .702. The Potts-Bennett Tests alone were thus almost as efficient as any combination studied.
5. The subtests of the Potts-Bennett and the MacQuarrie in general correlated less highly with the criterion than did the total scores on each test.
6. Although only percentile scores were available for the Bernreuter Personality Inventory, it appears to correlate with the criterion less well than the other tests.

Received May 23, 1945.

References

1. Bennett, Geo. K., and Gordon, H. Phoebe. Personality test scores and success in the field of nursing. *J. appl. Psychol.*, 1944, 28, 267-278.

2. Douglass, H. R., and Merrill, R. A. Prediction of success in the school of nursing. *Univ. Minnesota Stud. in the Prediction of Scholastic Achievement*, 1942, **2**, 17-31.
3. Garrison, K. C. The use of psychological tests in the selection of student nurses. *J. appl. Psychol.*, 1939, **23**, 461-472.
4. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill Book Co., Inc., 1936.
5. MacPhail, A. H., and Bernard, W. Ten years of intelligence testing. *Educ. & Psychol. Meas.*, 1943, **3**, 157-165.
6. Potts, Edith Margaret. Use of tests in selecting student nurses advantageous to hospital and student. *Hosp. Mgmt.*, 1941, **52**, 39-42.
7. Rainier, R. N., Rehfeld, F. W., and Madigan, M. E. The use of tests in guiding student nurses. *Amer. J. Nurs.*, 1942, **42**, 674-682.
8. Scruggs-Carruth, Margaret. *The predictive value of nursing school tests*. Unpublished Thesis: Southern Methodist University, Dallas, Texas. 1944.
9. Super, D. E. The Bernreuter Personality Inventory: A review of research. *Psychol. Bull.*, 1942, **39**, 94-125.
10. Tiffin, Joseph. *Industrial psychology*. New York: Prentice-Hall, Inc., 1942.
11. Williamson, E. G., Stover, R. D., and Fiss, C. B. The selection of student nurses. *J. appl. Psychol.*, 1938, **22**, 119-131.

Teachers College Students and the Minnesota Multiphasic Personality Inventory

Orpha Maust Lough

State Teachers College, Fredonia, New York

Between February 1944 and January 1945, 202 students at a New York State Teachers College took the Minnesota Multiphasic Personality Inventory in connection with the course in Child Development. The records of only the 185 unmarried, women students, the majority of whom were Freshmen when they took the Inventory, were used in this study as the number of men students was too small to be significant. Of these women students, 94 were enrolled in the General Curriculum and the remainder were enrolled in the Music Curriculum. Those enrolled in the General Curriculum were preparing to be elementary school teachers; those in the Music Curriculum, public school music teachers.

The purpose of this investigation was to determine: (1) if on the basis of the MMPI, there were any significant differences on any of the scales between those students enrolled in the Music Curriculum and those in the General Curriculum, or between these teachers college students and the general population as reported by the authors of the Inventory; (2) whether such an Inventory might be useful in the selection of students for admission to the teaching profession; (3) whether the Inventory indicates in these teachers college students the probabilities of developing those types of maladjustments which have been claimed to be predominant in various studies of school teachers.

The Minnesota Multiphasic Personality Inventory is a technique developed at the University of Minnesota and published in 1943. It is individually administered and consists of five hundred fifty statements in simple language, each printed on a separate card, covering a wide range of subjects including physical condition, morale, vocational interests, and social attitudes. The subject sorts these statements into three categories,—"True," "False," "Cannot Say." The decisions are recorded on a printed Record Sheet according to instructions given in the Manual, and scored on twelve different scales: three validating scales; The Question Score (?), The Lie Score (L), The Validity Score (F); and nine diagnostic categories: The Hypochondriasis Scale (H_s), The Depression Scale (D), The Hysteria Scale (H_y), The Psychopathic Deviate Scale (P_d), The

Interest Scale (M_t), The Paranoia Scale (P_a), The Psychasthenia Scale (P_t), The Schizophrenia Scale (S_c), and The Hypomania Scale (M_a).

Findings

The ages of the subjects in this study ranged from sixteen years to twenty-three years, with a mean age of 18.8 years. The mean age of those in the General Curriculum was 19.0 and those in the Music Curriculum, 18.6. These students were given the Revised Alpha Examination Form 5. The mean percentile rank on the Alpha for both groups taken together was 94.0; for the General Curriculum 93.2 and for the

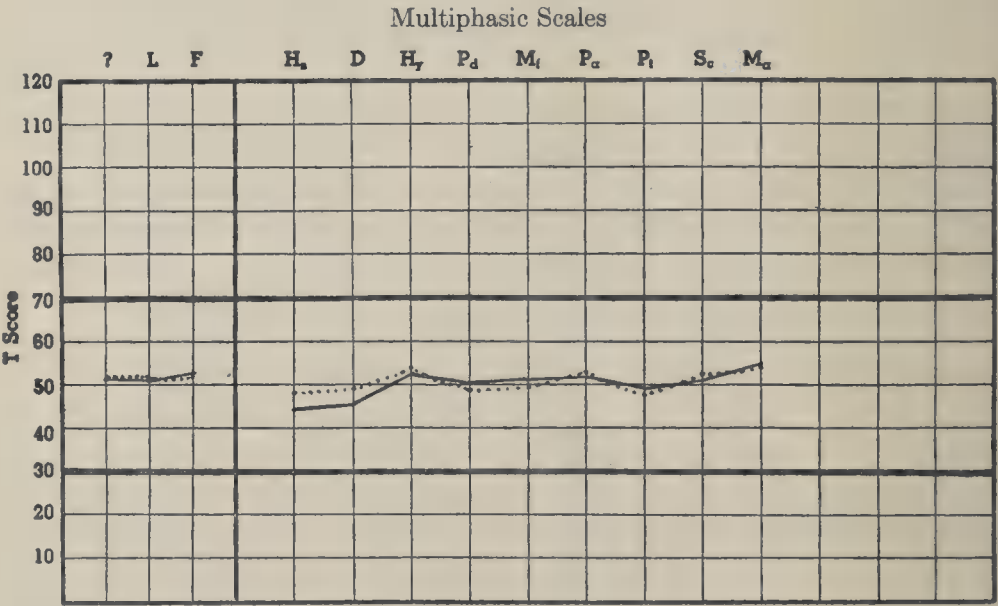


FIG. 1. T-score profiles on the Minnesota Multiphasic Personality Inventory for Music Students and General Students enrolled in a New York State Teachers College. Note: solid line = 74 General Students; dotted line = 111 Music Students.

Music Curriculum 94.4. Neither of these differences between the means is statistically significant (see Table 1).

Figure 1 shows graphically the profiles for the music students and the general students based on the means of the T-scores for each scale.

Both of the profiles approach a fairly straight line at the T-score mean level of 50 although they are a little higher generally than the one reported by Schmidt (15) for 98 normal men. The profiles tend to be fairly similar although some differences may be noted. The means of the music students on the psychasthenia scale is lower than that of the general students. The lowest point on the profile of the music students is the hypochondriasis scale and the highest point, the hypomania scale. The

means of the general students on the hysteria scale and depression scale is lower than those of the music students. The lowest point in the profile of the general students is the psychasthenia scale and the highest is the hypomania scale. Both groups are approximately one-half a standard deviation above the mean level on the hypomania scale. Schmidt (15) found a decrease on the hypomania scale for all profiles of both the normal and clinical groups of men.

Table 1

Comparison of the Teachers College Students Enrolled in the Music Curriculum and in the General Curriculum on the Separate Scales of the Minnesota Multiphasic Personality Inventory

	Gen. Curr. N = 74			Music Curr. N = 111			Critical Ratio Between Gen. Curr. and Music Curr.	All Students N = 185	
	Range	Mean M ₁	S.D. M ₁	Range	Mean M ₂	S.D. M ₂		Mean	S.D.
Age	22-5 to 17-3	19.0	1.33	23-1 to 16-8	18.6	1.22	.21	18.8	1.26
Rev. Alpha Form 5 M.M.P.I.	71-99	93.2	6.10	76-99	94.4	4.84	.16	94.0	5.33
?	50-66	51.3	3.10	50-66	51.3	3.16	.008	51.3	3.14
L	50-66	51.7	3.24	50-66	51.7	3.86	.000	51.7	3.63
F	50-73	52.7	5.02	50-70	52.6	4.92	.018	52.6	4.97
H _a	37-67	45.7	7.91	37-79	48.1	8.02	.210	47.1	7.98
D	28-73	47.1	12.06	32-71	49.3	8.53	.150	48.4	9.42
H _y	36-80	52.5	9.03	33-77	52.9	8.72	.030	52.7	9.05
P _d	35-91	50.4	12.87	35-82	49.3	9.85	.071	49.7	11.16
M _t	30-74	50.9	9.25	26-74	49.9	9.25	.076	50.3	9.25
P _a	33-79	51.7	8.46	33-82	52.5	8.84	.072	52.2	8.69
P _t	34-90	49.9	8.90	36-75	47.9	9.86	.150	48.7	9.21
S _c	39-83	50.7	9.32	37-74	51.8	9.23	.080	51.4	9.27
M _a	37-72	54.6	8.31	30-84	54.5	11.22	.004	54.6	10.14

Table 1 presents the ranges, means, standard deviations for each group and for all the students, together with the difference of the means, the standard errors of the difference, and the critical ratios between the two groups.

From the data given in Table 1, it is apparent that there is no significant difference on any of the scales between the students enrolled in the two curricula as none of the critical ratios is three or more.

According to the manual of instruction accompanying the MMPI, normal persons do not often score above 70; but if the environmental pressure is small, or if other personality factors are favorable, a person may score over 70 and yet escape need for special attention. Table 2 shows the percentage of these students with T-scores over 70 on the separate scales of the Inventory.

Table 2

Percentage of the Teachers College Students with T-scores above 70 on the Separate Scales of the Minnesota Multiphasic Personality Inventory

Scale M.M.P.I.	Per cent of 74 Gen. Stud. with T-scores above 70	Per cent of 111 Music Stud. with T-scores above 70	Per cent of 185 Students with T-scores above 70
?	0	0	0
L	0	0	0
F	1.3	0	.5
H _s	0	1.8	1.1
D	2.6	.9	1.6
H _y	4.0	5.4	4.9
P _d	6.5	2.7	3.8
M _f	4.0	1.8	2.7
P _a	5.4	4.5	4.9
P _t	1.3	2.7	2.2
S _c	5.4	3.6	3.8
M _a	4.0	10.0	7.6

On the basis of the percentage scoring above 70 on the various scales of the Inventory as given in Table 2, it would appear that there may be some personality differences between the students in the two curricula. The extremely high scores among the general students were on the psychopathic-paranoia-schizophrenia scales while among the music students, the extremely high scores were on the hypomania-hysteria-paranoia scales. When the two groups were combined, the highest percentage of the T-scores over 70 were on the hypomania scales and in decreasing order of percentage, on the hysteria, paranoia, psychopathic deviate and schizophrenia scales.

From these data it would appear that the teachers college students who took the MMPI were on the whole normal and stable. They show some slight disposition toward hypomania which is characterized by over-productivity in thought and action; ambition, vigor; activity and enthusiasm, although somewhat depressed at times; inclination to undertake too many things at a time, to stir up projects and then lose interest in them; a disposition to disregard social conventions. There are ap-

parently no very significant differences between those enrolled in the music and in the general curricula except among those with T-scores over 70.

It may be that the slight tendency toward hypomania is characteristic of the late adolescent or young adult. It may be that this tendency is characteristic of women since Landis and Page (10) report that the incidence of manic-depressive psychoses among men is but four-fifths as high as among women, while the incidence rate of schizophrenia is 15% higher among men than among women.

In a study of seven hundred maladjusted school teachers who were or had been in hospitals for the mentally ill, Mason (12) found that the most common diagnosis among this group was dementia praecox and the second highest was manic-depressive psychosis. Thirty-seven per cent of the group suffered from schizophrenia, 31% of the men and 40% of the women. Those diagnosed as suffering from manic-depressive psychosis constituted 24% of the group, 19% among the males and 26% among the females. The figures of New York state as a whole confirmed these findings that the largest percentage of teachers committed to state hospitals suffered from dementia praecox (27%) and the next largest clinical group was manic-depressive (14%). As shown by the present data, this group of women students preparing to be teachers showed little or no indication of schizophrenic trend while there was some slight evidence of manic-depressive tendencies.

In a study by Harmon and Weiner (2) on the use of the MMPI in vocational advisement of disabled veterans they state that "elevation on Psychopathic or Hypomania scales indicates the type of personality most likely to adjust in jobs of a relatively undisciplined nature, where individual initiative and aggressiveness are at a premium, and which afford a maximum of variety in work processes, locale, or associates." In the light of this statement it would appear that teaching may be a good vocation for those with hypomanic trends in as much as teachers who use modern methods in their work with children are continuously varying their work processes, need initiative and aggressiveness, and their work is fairly unregimented. On the other hand, a teacher who is even in the early stages of manic-depressive psychosis may have a great disorganizing effect upon the children under her direction because of her instability, erratic behavior, and lack of wholesome, well-integrated personality.

Summary

It would appear from this study of 185 unmarried women students in a teachers college that, on the basis of the Minnesota Multiphasic Personality Inventory:

1. They are a relatively stable, normal group with a very slight tendency toward hypomania.

2. There are no significant differences between those preparing to teach in the elementary grades and those preparing to become public school music teachers.

3. There may be some slight relationship between the hypomania trend found in these students and the large incidence of manic-depressive psychosis found among teachers who have been hospitalized. These students show no trends toward schizophrenia which was found to be the most common diagnosis among a group of seven hundred hospitalized teachers (12). As age appears to be a factor in the onset of psychoses (10), the fact that students now entering New York state teachers colleges are required to complete a four-year curriculum, may eliminate some of those inclined toward schizophrenia but probably not those with a predisposition to manic-depressive psychosis. Such trends, however, might be among the many factors which contribute to the elimination of students between the Freshman and Senior years and between graduation and status as a classroom teacher. A follow-up study should be made in order to verify this conclusion.

4. It may be, on the basis of this study together with that of Harmon and Weiner (2), that certain scales of the MMPI might be useful as one of the instruments in the selection of students for admission to the teaching profession. This research may provide normative data for guidance workers who are counseling incoming freshmen in other teachers colleges. However, a great deal more research on the personality characteristics of prospective teachers needs to be done before such a program is instituted.

Received June 13, 1945.

References

1. Altus, W. D., and Bell, H. M. The validity measures of certain maladjustment in an army special training center. *Psychol. Bull.*, 1945, 42, 98-103.
2. Harmon, L. R., and Wiener, D. N. Use of the Minnesota multiphasic personality inventory in vocational advisement. *J. appl. Psychol.*, 1945, 29, 132-141.
3. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule: I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249-254.
4. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule: II. A differential study of hypochondriasis. *J. Psychol.*, 1940, 10, 255-268.
5. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule: III. The measurement of symptomatic depression. *J. Psychol.*, 1942, 14, 73-84.
6. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule: IV. Psychasthenia. *J. appl. Psychol.*, 1942, 26, 614-624.
7. McKinley, J. C., and Hathaway, S. R. The Minnesota multiphasic personality inventory: V. Hysteria, hypomania and psychopathic deviate. *J. appl. Psychol.*, 1944, 28, 153-174.

8. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota multiphasic personality inventory*. Minneapolis: University of Minnesota Press, 1943; more recently, New York: The Psychological Corporation.
9. Hathaway, S. R. The personality inventory as an aid in the diagnosis of psychopathic inferiors. *J. consult. Psychol.*, 1939, 3, 112-117.
10. Landis, C., and Page, J. *Modern society and mental disease*. New York: Farrar and Rinehart, 1938, 39-40.
11. Leverenz, Major C. S. Minnesota multiphasic personality inventory: An evaluation of its usefulness in the psychiatric service of a station hospital. *War Med.*, 1943, 4, 618-629.
12. Mason, F. V. A study of seven hundred maladjusted school teachers. *Mental Hygiene*, 1931, 15, 576-599.
13. McKinley, J. C., and Hathaway, S. R. The identification and measurement of the psychoneurosis in medical practice: The Minnesota multiphasic personality inventory. *J. Am. med. Ass.*, 1943, 122, 161-167.
14. Schiele, B. C., Baker, A. B., and Hathaway, S. R. The Minnesota multiphasic personality inventory. *Journal-Lancet*, 1943, 63, 292-297.
15. Schmidt, H. O. Test profiles as a diagnostic aid: The Minnesota multiphasic inventory. *J. appl. Psychol.*, 1945, 29, 115-131.

The Occupational Adjustment Characteristics of a Group of Sexually Promiscuous and Venereally Infected Females *

Robert D. Weitz

Jersey City, New Jersey

This study is the second of a series in which the writer is engaged in connection with a program of social and vocational rehabilitation of sexually promiscuous and venereally infected females treated at the Midwestern Medical Center,—an intensive treatment center of the United States Public Health Service.

It was pointed out in the first study (9), based on 500 cases, that the patients treated were generally below normal intelligence, the majority falling in the defective and borderline defective mental range. It was further pointed out that the group as a whole was more than a full year retarded scholastically, the eighth grade being the median level completed.

This study is concerned with the occupational adjustment which generally characterized the group. As a basis of comparison, the data obtained were compared with those of a corresponding age group of unselected female job applicants who applied for work through the United States Employment Service (USES) in St. Louis during the years of 1941 and 1942.

The Problem

By virtue of the fact that many writers have dealt with the subject of work as a socio-economic factor in the incidence of venereal disease, it was the purpose of this study to ascertain whether there were reliable differences in the occupational adjustments of a group of sexually promiscuous female patients treated at the Midwestern Medical Center and a group of ostensibly normal girls. Occupational adjustment was considered in terms of two factors: (1) the nature of the job, as indicated by occupational title ¹ and (2) job stability—as measured by the longest consecutive number of months that each subject was gainfully employed by a single employer.

* This report is based on a research study conducted by the writer while affiliated with the United States Public Health Service. Appreciation is extended to Virginia S. Lenobel, psychological assistant, for her assistance in the testing program.

¹ The occupational titles assigned to the subjects of both groups were based on the principles of job classification embodied in the Dictionary of Occupational Titles (4).

Related Studies

The literature of the earlier studies of wayward females in general indicates again and again evidence of vocational maladjustment and the need of guidance.

In a study of 181 delinquent females who had been committed to the Illinois State Training School for Girls, Abbott and Breckenridge (1) found that relatively few of the subjects had worked at or had prepared for jobs requiring skill or training. Domestic work, waitress work and other unskilled jobs characterized the usual type of employment. It was also observed that frequent change of job was commonplace in the work histories of these girls. Broughton (2), in discussing the relationships between jobs and prostitution pointed out that the girls who headed for the "oldest profession" have had few opportunities for decent jobs. He stressed the need for vocational guidance and vocational education in the redirection of the girls who might become the next crop of prostitutes. In the same vein, Parran (6), commenting on the reduction of prostitutes in Russia, pointed out that the girls apprehended there are given trade training as well as medical treatment to aid in their reclamation.

In a study reported by De La Caro (3), of a group of adolescent girls, mostly prostitutes, interned in Caguas, Puerto Rico, the need for vocational orientation was again pointed out. De La Caro stated:

"The facts obtained from this study prove that a large proportion of these girls are in a favorable condition for a possible rehabilitation and that a coordinated program of services could save them. The venereal disease hospitals have already made provisions for a program of recreation and vocational orientation in the hospital, but this is not sufficient. The period of hospitalization of these patients is generally much too short to assure their permanent social rehabilitation. We believe it is the responsibility of the community to continue this work."

Ness (5), in his talk at the Puerto Rico Regional Conference on Social Hygiene, held February 1944, clearly indicated that vocational reorientation was highly necessary as part of a more general social rehabilitation of sexually promiscuous females treated for venereal disease. He called attention to the "revolving door" theory, i.e., letting the girls go back, at the completion of their medical treatment, to the same conditions that helped produce their maladjustment and disease.

In a recent study by Rachlin (7) of a mixed group of colored and white Midwestern Medical Center patients who had been treated prior to the group included in the present study, he found that it was difficult to determine their earning power, because the subjects had not worked steadily. In listing the jobs held by the girls, he showed that waitress, factory laborer, housekeeper and other unskilled jobs were most common to the group.

Like the authors of the other studies referred to above, Rachlin maintains that venereal disease can be eradicated only by a closely coordinated medical and social rehabilitation program. It is in this vein that he recognizes the need for vocational guidance and reorientation.

The Subjects

The group which served as the basis for comparison was comprised of 225 female job applicants, all residents of St. Louis, who had filed job applications with the USES. Their chronological ages ranged from 17 to 30 years inclusive with a mean level of 22.9 years. A similar number of cases was selected from the original group of the 500 sexually promiscuous and venereally infected females treated at the Midwestern Medical Center to match as closely as possible the age distributions of the USES group. These cases, too, were all residents of St. Louis. In order to obtain the 225 cases for the hospital group, it was necessary to extend their age range to 35 years. The mean age level for the cases included in

Table 1

The Comparative Age Distributions of the Midwestern Medical Center Patients and the USES Job Applicants Included in the Study

Age in Years	Hospital Cases		USES Cases	
	No.	%	No.	%
35-35.9	1	.44		
34-34.9	2	.89		
33-33.9	3	1.33		
32-32.9	3	1.33		
31-31.9	3	1.33		
30-30.9	4	1.78	13	5.78
29-29.9	6	2.67	11	4.89
28-28.9	6	2.67	7	3.11
27-27.9	8	3.56	12	5.33
26-26.9	9	4.00	17	7.56
25-25.9	8	3.56	10	4.44
24-24.9	13	5.78	12	5.33
23-23.9	9	4.00	9	4.00
22-22.9	34	15.12	14	6.22
21-21.9	25	11.11	27	12.00
20-20.9	29	12.89	33	14.67
19-19.9	22	9.78	20	8.89
18-18.9	38	16.88	38	16.89
17-17.9	2	.89	2	.89
Totals	225	100.00	225	100.00
Mean		22.8		22.9
S.D.		3.99		3.83

Table 2

The Reliability of the Differences Between the Groups in Chronological Age and Duration of Longest Job

	Hospital Group		USES Group		D	Diff.	$\frac{D}{\sigma \text{ diff.}}$
	M	S.D.	M	S.D.			
Chronological Age in Years	22.8	3.99	22.9	3.83	.17	.37	.46
Longest Job in Months	15.1	17.01	31.4	33.22	16.31	2.88	5.66

this group was 22.8 years, virtually the same mean as found for the USES cases.

To reduce the influence in job adjustment which might be attributed to race differences, only white girls were included in this study.

The Findings

That the hospital group and the USES group were closely matched for age is seen in Table 1. The mean age for the hospital cases was 22.8 years, with a standard deviation of 3.99; whereas, the USES job applicants showed a mean age of 22.9 years and a standard deviation of

Table 3

A Comparison of the Classified and Unclassified Subjects Found in Both Groups *

Subjects	Hospital Group		USES Group	
	No.	%	No.	%
Classified	214	95.0	158	70.0
Unclassified	11	5.0	67	30.0
Totals	225	100.0	225	100.0

* The subjects with one month or more of gainful experience in a single occupation were designated as *classified*; those with less than one month were designated as *unclassified*.

3.83. As seen in Table 2, the critical ratio ² determined for these age differences (CR = .46) indicates that there was no significant difference between the mean age of the matched groups.

Table 3 reveals that 214 of the 225 hospital patients studied, or 95 per cent, had experience of at least one or more months in a single job;

² A significant difference between groups requires a critical ratio ($\frac{D}{\sigma \text{ diff.}}$) of 3.00 or higher (8).

whereas, for the 225 USES job applicants it is seen that only 158 of the 225 cases studied, or 70 per cent, had work experience of a month or more on a single job. This difference was due to the fact that the USES group included several girls who were recently out of school and were entering the field of work for the first time.

Comparing the subjects of both groups wherein job titles were designated (i.e., where the individual had one month or more of gainful experience in a single job) it is seen in Table 4 that the mean work period

Table 4
A Comparison of the Longest Work Periods for the Occupationally
Classified Subjects

Longest Work Period in Months	Hospital Cases		USES Cases	
	No.	%	No.	%
151-160			2	1.27
141-150			2	1.27
131-140			0	.00
121-130			1	.63
111-120	1	.47	3	1.90
101-110	1	.47	1	.63
91-100	2	.93	3	1.90
81- 90	1	.47	3	1.90
71- 80	0	.00	4	2.53
61- 70	0	.00	2	1.27
51- 60	2	.93	9	5.70
41- 50	6	2.80	8	5.06
31- 40	7	3.27	14	8.86
21- 30	23	10.75	28	17.72
11- 20	46	21.50	22	13.92
1- 10	125	58.41	56	35.44
Totals	214	100.00	158	100.00
Mean		15.1		31.4
S.D.		17.0		33.2

for the hospital group was 15.1 months, with a standard deviation of 17.01. For the USES group the mean work period was 31.4 months, with a standard deviation of 33.22. The critical ratio as shown in Table 2, for these cases was 5.66, indicating a statistical significant difference.

A comparison of the groups based on the occupational classifications assigned to the subjects is shown in Table 5. Outstanding among the findings revealed in this table are the following: (1) the complete absence of the hospital patients on the professional and managerial level, (2) their comparatively poor representation in the skilled and semi-skilled

Table 5

The Comparative Distributions of the Occupational Classifications of the Groups

Occupational Classification	Hospital Cases		USES Cases	
	No.	%	No.	%
Professional and Managerial	0	.00	5	3.16
Clerical and Sales	16	7.48	15	9.49
Service	93	43.46	11	6.96
Agricultural, fishery, forestry, etc.	2	.93	0	.00
Skilled	21	9.81	33	20.89
Semi-skilled	42	19.62	57	36.08
Unskilled	40	18.70	37	23.42
Totals	214	100.00	158	100.00

trades and (3) the preponderance of the hospital patients in the service occupations as, for example, waitresses, houseworkers, bar-maids, etc.

Summary and Conclusions

On the basis of the findings of this study, it would appear that the sexually promiscuous and venereally infected patients who comprise the Midwestern Medical Center group differ from a group of ostensibly normal girls with reference to work adjustment. This is seen in the following:

1. From the standpoint of length of service for a single employer, the hospital cases were not as stable as the USES group, the latter persons having worked twice as long on the average.
2. The hospital patients were found much more frequently in jobs requiring little or no skill and training. This is evidenced by their complete absence on the professional and managerial level, their comparatively poor representation in the skilled and semi-skilled trades and their relatively great distribution in the service occupations.

No doubt there are many reasons why the girls, who are known to be sexually promiscuous, tend to manifest a generally inferior work adjustment. To be sure, low intelligence, emotional instability, economic insecurity, or various combinations of these as well as other factors, may be included as basic causes in the total maladjustment syndrome of the sexually promiscuous female, but it is not within the scope of this paper to determine the causes of the maladjustment. It is rather the purpose here mainly to recognize that these individuals are different and are, consequently, in need of aid.

Because of the fact that work plays so great a role in everyday life,

it is important that job adjustment be of prime consideration in the rehabilitation of the individual. An adequate vocational guidance program is therefore a necessary adjunct to any program, medical or otherwise, concerned with social reclamation. The role of the vocational counselor will, of course, be determined by the caseholding policy under which the institution operates. Where patients are held for medical treatment over a period long enough to permit adequate guidance work-up, including aptitude testing, the counselor alone can do a great deal toward the reorientation of the patients. On the other hand, where the latest intensive methods of therapy are used, and the patient turnover is rapid, the counselor can serve best by screening the patients and referring them to such community agencies as are available and willing to cooperate in the rehabilitation program.

Received April 19, 1945.

References

1. Abbott, Grace, and Breckenridge, Sophonisba. *The delinquent child*. New York: Russel Sage Foundation, 1912. Pp. 76.
2. Broughton, Philip S. Prostitution and the War. *Public Affairs Pamphlet*, No. 65, 1942. Pp. 24.
3. De la Caro, Dolores G. Youth in crisis: New horizons for our girls in trouble. *J. soc. Hyg.*, 1944, 30, 244-249.
4. *Dictionary of Occupational Titles*, Part I. Washington: U. S. Government Printing Office, 1939.
5. Ness, Eliot. Social protection in venereal disease control. *J. soc. Hyg.*, 1944, 30, 226-231.
6. Parran, Thomas. *Shadow on the land, syphilis*. New York: Reynal and Hitchcock, 1937.
7. Rachlin, H. L. A sociologic analysis of 304 female patients admitted to the Midwestern Medical Center, St. Louis, Mo., *Venereal Disease Information*, 1944, 25, 265-271.
8. Sorenson, Herbert. *Statistics for students of psychology and education*. New York: McGraw-Hill Book Company, Inc., 1936.
9. Weitz, Robert D., and Rachlin, H. L. The mental ability and educational attainment of five hundred venereally infected females. *J. soc. Hyg.*, 1945, 31, 300-303.

The Effect of Prolonged Mild Anoxia on Speech Intelligibility *

G. M. Smith

College of the City of New York

In two earlier papers in collaboration with C. P. Seitz (3, 4), it was demonstrated that speech intelligibility suffers a statistically reliable decrement at simulated altitudes as low as 16,900 ft., under certain conditions of initial difficulty. The decrement is a function of the initial intensity of the stimulus sounds, being greater for lower intensities. In view of the common practice in long-range bombing of flying to the region of the target at comparatively low altitudes without the use of oxygen masks, it was felt that an investigation of possible hearing losses under the stress of mild but prolonged anoxia might be of some interest. In the present study tests of speech intelligibility were made at intervals throughout a number of eight-hour sessions in a nitrogen dilution chamber in which an altitude of approximately 10,000 ft. was simulated.¹

Method and Materials

The method of observing the effect of oxygen deprivation on the subjects' ability to perceive speech sounds and the test materials employed were the same as those employed in the second study mentioned above (4). The subjects indicated their responses to the stimulus words on check lists. The stimulus materials were made up in part from standard word lists developed by the Bell Telephone Laboratories, covering the more frequent sounds that occur in common speech. Other lists were constructed on the principles employed by the Bell Laboratories (2). Intelligibility for vowel sounds was tested by lists of monosyllables all having the same initial and final consonants in any one list; e.g., *suit, sit, sat, set*, etc. Consonant intelligibility was tested by similar lists, each involving a constant vowel sound, but a variation in the initial or final consonant; e.g., *nor, bore, yore, more*, etc.

To insure uniformity of stimuli the test items were recorded by means

* I am indebted to the Linde Air Products Co. for a liberal grant of oxygen and nitrogen, and to Messrs. Mortimer Feinberg and Max Rosenbaum for valued clerical and statistical assistance.

¹ The effects of the experimental variable on several other functions studied during the same sessions are reported elsewhere (5, 6, 7).

of the high fidelity equipment at the National Broadcasting Company studios in New York City.² To minimize the effect of wear these recordings were later put in semi-permanent form by the RCA Manufacturing Co. at Camden, N. J.³ The recordings were played back through a Fairchild pick-up⁴ coupled with a Presto amplifier⁵ (especially adapted so as to give a relatively flat response curve) and Western Electric ear-phones.⁶

The experiments were carried out in the nitrogen dilution chamber of the College of the City of New York, described in a previous paper (4). This maintains temperature and humidity at constant and comfortable levels and provides quite adequate sound-proofing. The simulated altitudes for the four experimental runs averaged 9,993 ft. (corresponding to an oxygen percentage of 14.3). The average altitudes maintained during the individual runs were 8,930 ft., 10,610 ft., 9,810 ft., and 10,620 ft. Samples of the chamber air were taken at intervals of approximately one hour on every run and were analyzed by means of the Haldane-Henderson-Bailey gas analysis apparatus. The mean deviation from the general average of the 38 individual samples analyzed was 680 ft. On the four control runs there was, unfortunately, some over-compensation for the leakage from the experimenter's mask, which caused a moderate climb from sea level during the runs, the average altitude for all four control runs being 1,810 ft. The mean deviation of the individual readings from this average was 1,510 ft. It is quite improbable that an altitude of this order, especially one simulated by reduced oxygen tension without change in total pressure, could have an adverse effect on hearing. Furthermore, the data from the four control runs indicated that on the runs for which the altitude was nearer to sea level the performance of the subjects was generally worse rather than better. It therefore seems justifiable to regard any altitude effects obtained on the intentional altitude runs as applicable to the 10,000 ft. level.

The carbon dioxide problem was effectively solved by means of an air conditioning device which circulated the chamber air through tubes of Shell Natron under forced draft. The 65 individual CO₂ readings, for both the altitude and control runs combined, averaged 0.28%, with a mean deviation of 0.10%.

² I am indebted to the National Broadcasting Co. and to Mr. R. A. Lynn of the Engineering Dept. for their cooperation.

³ I am indebted to the RCA Manufacturing Co. and to Mr. W. L. Tesch of the Record Engineering Dept. for their courtesies.

⁴ Turntable unit model 199 and pick-up model 209.

⁵ Model 87B.

⁶ Type 588A.

Procedure

Twelve male college students, including several ASTP volunteers⁷ served as subjects. Their ages ranged from 17–20 years, with the median at 18 years. They worked in four groups of three, each group being tested both at the experimental altitude of approximately 10,000 ft. and at the control altitude of approximately 1,800 ft. for a continuous eight-hour run. The order of the altitude and control runs was reversed for alternate groups of subjects so as to minimize the effect of practice. The two runs were separated by an interval of one week. The usual precautions to allay fear of the chamber were taken, the experimenter remaining in the chamber throughout the run. To keep the suggestion factor constant for the two runs in which each subject participated, the procedure of the experimental run, including the manipulation of oxygen and nitrogen valves and the wearing of an oxygen mask by the experimenter, was carefully imitated on the control runs. Judging by the subjects' reactions, the deception was quite general. To relieve the tedium the subjects were permitted to read or study quietly when not engaged in taking tests. During the testing the subjects were comfortably seated and were equipped with standard Western Electric earphones such as are used by the American Airlines on transport planes.⁸

Record booklets made up of four separate tests, each containing check rows for 11 vowels and 24 consonants, were employed. This made a total of 44 vowel items and 96 consonant items. The order of the four tests was varied from one testing period to the next in order to minimize practice effects. The testing periods, which were approximately 20 minutes in duration, came after average elapsed times in the chamber of $\frac{3}{4}$ hr., $2\frac{1}{4}$ hrs., $4\frac{3}{4}$ hrs., and $6\frac{3}{4}$ hrs., approximately. Between the second and third testing periods a high protein standardized lunch intervened.⁹ This began after an elapsed time of approximately $3\frac{3}{4}$ hrs. and was finished within $\frac{1}{2}$ hr.

The sound level of the stimulus words was the same as that used in the second study in collaboration with C. P. Seitz (4). This was set inten-

⁷ I am indebted to Colonel Raymond P. Cook for his cooperation in permitting the use of Army volunteers for this rather tedious ordeal.

⁸ Type 588A. The response curves of one set of phones was tested by the Stevens Institute of Technology through the courtesy of President H. N. Davis and Dr. H. Burris-Meyer. Though the curves are peaked rather sharply at 1,000 cps., they are relatively flat (± 10 db.) between 2,000 and 8,000 cps. The transmission system as a whole was sufficiently free from distortion to make possible a ready identification of the speakers' voices. I am indebted to the American Airlines and to Messrs. D. W. Rentzel and H. A. Wolfe for the use of the phone sets.

⁹ Two ham or cheese sandwiches and one pint of milk. For each subject the diet was the same during the experimental and control runs.

tionally at a fairly low value ¹⁰ so that the effect of anoxia, if any, might be more readily demonstrated. Specifically, the mean consonant articulation value (the per cent of correct responses) was in the range 50-60%, the exact value varying somewhat with the sample of subjects used. At this sound level, and with the same test materials and reproducing equipment as those employed in the present study, there was no reliable decrement in intelligibility observed after an exposure to a simulated altitude of 13,600 ft. for approximately one hour. However, in this earlier study a simulated altitude of 16,900 ft. did produce a reliable decrement.

Results

The principal data are summarized in Table 1, which gives the articulation values (the per cent of correct responses) for vowels, consonants, and standard syllables for each of the twelve subjects, for both the altitude and the control runs, for each of the four testing periods. Standard syllable articulation values were calculated from the vowel and consonant values by the Fletcher and Steinberg formula $S = 1 - (1 - VC^2)^{0.9}$ derived empirically from extensive observations in the Bell Laboratories (2). The means and the probable errors of these means also appear in this table. As is to be expected, the articulation values for vowels are consistently better than those for consonants for all subjects under both altitude and control conditions. For all four periods the mean performances under the control conditions are superior to the mean performances at altitude for each of the three criteria. The standard syllable criterion is the most meaningful since it takes into account both vowel and consonant articulation and most nearly approximates speech. The mean syllable articulation values for both the altitude and the control runs are plotted in Figure 1. This gives us an impression of maximum altitude handicap at periods II and III, 2¼ hrs. and 4¾ hrs. after the beginning of the run, respectively, and also suggests a marked end-spurt at the 6¾ hr. mark.

As a check on the impression given by this figure, Fisher's *t*-statistic was calculated for each of the four periods to see whether the apparent differences were reliable. The results of these calculations are presented in Table 2. This indicates that the small differences which appear between the altitude and the control performances at the ¾ hr. and the 6¾ hr. periods are quite unreliable. However, the differences at the 2¼ hr. and 4¾ hr. periods, though not strictly reliable by rigorous stand-

¹⁰ Approximately 24 db. through the phones, against an external background sound level of approximately 70 db. in the chamber due to fans in temperature control and circulation systems, primarily.

Table 1

Articulation Values for Vowels, Consonants, and Standard Syllables at Four Different Periods

Sub- ject	Period I					
	Elapsed Time: $\frac{3}{4}$ Hr.					
	Control			Altitude		
	Vowel* %	Conso- nant† %	Syllable‡ %	Vowel* %	Conso- nant† %	Syllable‡ %
1	75.0	47.0	15.0	75.0	34.5	9.0
2	100.0	65.5	39.5	95.5	44.0	17.0
3	97.5	59.5	32.0	93.0	45.0	17.0
4	95.5	73.0	47.5	95.5	56.5	28.0
5	100.0	64.5	38.5	47.5	19.0	1.5
6	100.0	78.0	57.0	97.5	67.5	41.0
7	97.5	63.5	36.5	97.5	86.5	69.5
8	100.0	56.5	29.5	100.0	72.0	48.0
9	95.5	56.5	28.0	97.5	57.5	29.5
10	88.5	50.0	20.0	100.0	62.5	36.0
11	70.5	37.5	9.0	95.5	55.0	26.5
12	97.5	47.0	19.5	95.5	49.0	21.0
Mean	93.1	58.2	31.0	90.8	54.1	28.7
P.E.M	1.86	2.23	2.64	1.38	3.21	3.29
Sub- ject	Period II					
	Elapsed Time: 2 $\frac{1}{4}$ Hrs.					
	Control			Altitude		
	Vowel %	Conso- nant %	Syllable %	Vowel %	Conso- nant %	Syllable %
1	86.5	47.0	17.5	86.5	35.5	10.0
2	100.0	73.0	49.5	97.5	55.0	27.5
3	100.0	62.5	36.0	93.0	49.0	20.5
4	95.5	67.5	40.0	66.0	38.5	9.0
5	95.5	70.0	43.5	95.5	58.5	30.0
6	100.0	81.5	62.5	100.0	66.5	41.0
7	100.0	79.0	58.5	100.0	75.0	52.5
8	100.0	62.5	36.0	100.0	69.0	44.0
9	82.0	54.0	20.0	100.0	56.5	29.5
10	97.5	36.5	11.5	100.0	52.0	25.0
11	93.0	41.5	14.5	88.5	48.0	18.5
12	97.5	49.0	21.5	88.5	54.0	23.5
Mean	95.6	60.3	34.2	92.9	54.8	27.6
P.E.M	1.05	3.01	3.46	1.72	2.10	2.40

Table 1—*Continued*

Period III						
Elapsed Time: 4½ Hrs.						
Sub- ject	Control			Altitude		
	Vowel %	Conso- nant %	Syllable %	Vowel %	Conso- nant %	Syllable %
1	79.5	40.5	11.5	79.5	33.5	8.0
2	97.5	72.0	47.0	100.0	51.0	24.0
3	100.0	62.0	35.5	97.5	46.0	20.0
4	100.0	70.0	45.5	95.5	55.5	27.0
5	100.0	73.0	49.5	97.5	67.5	41.0
6	100.0	73.0	49.5	95.5	74.0	48.5
7	100.0	74.0	51.5	100.0	75.0	52.5
8	100.0	61.5	35.0	93.0	73.0	46.0
9	97.5	44.0	17.5	100.0	47.0	20.5
10	97.5	45.0	18.0	95.5	36.5	11.5
11	88.5	48.0	18.5	88.5	44.0	15.5
12	95.5	54.0	25.5	97.5	51.0	23.0
Mean	96.3	59.7	33.7	95.0	54.5	28.1
P.E. _M	1.09	2.73	3.14	0.98	2.95	3.06
Period IV						
Elapsed Time: 6½ Hrs.						
Sub- ject	Control			Altitude		
	Vowel %	Conso- nant %	Syllable %	Vowel %	Conso- nant %	Syllable %
1	88.5	55.0	24.5	79.5	41.5	12.0
2	97.5	69.0	43.0	95.5	65.5	38.0
3	100.0	64.5	37.0	95.5	63.5	35.5
4	100.0	76.0	54.5	95.5	61.5	33.0
5	100.0	72.0	48.5	97.5	61.5	34.0
6	100.0	78.0	57.0	100.0	66.5	41.0
7	100.0	67.5	42.5	100.0	82.5	64.0
8	97.5	44.0	17.0	100.0	70.0	45.5
9	97.5	56.5	28.5	86.5	60.5	29.0
10	100.0	45.0	18.5	75.0	38.5	10.5
11	100.0	44.0	17.5	75.0	49.0	17.0
12	91.0	51.0	21.5	97.5	60.5	33.0
Mean	97.7	60.2	34.2	91.5	60.1	32.7
P.E. _M	0.65	2.67	3.15	2.03	2.08	2.71

* There were 44 vowel items.

† There were 96 consonant items.

‡ $S = 1 - (1 - VC^2)^{0.9}$.

ards, are at least suggestive. The probabilities that the differences could have arisen on the basis of chance are .14 and .07., respectively.¹¹ These more nearly reliable values correspond to the higher percentages obtained when the ratios of control to altitude performances are calculated. The control performance is 24% better than the altitude performance after a 2¼ hr. exposure and 20% better after a 4¾ hr. exposure; whereas the control performance is only 8% better at the ¾ hr. point and 5% better at the 6¾ hr. point. The relatively poor performance on

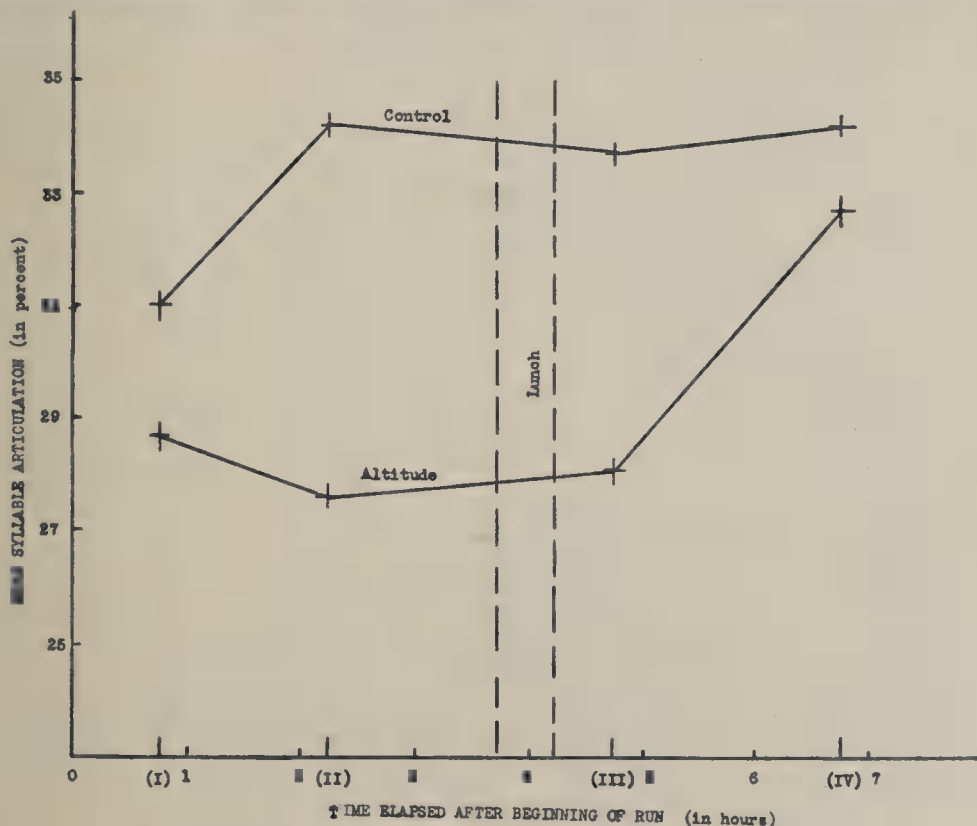


FIG. 1. Mean syllable articulation at the four test periods.

the control run at the first test period may quite possibly be the result of the belief on the part of the subjects that they were at altitude.

The question of the end-spurt in the performance at altitude at the 6¾ hr. period deserves some consideration. This is so marked as to practically wipe out any indication of reduced capacity to perceive speech sounds at the 10,000 ft. altitude. It strongly suggests that the apparent altitude effect shown at periods II and III is not primarily an auditory

¹¹ When these probabilities are combined by Fisher's technique for combining probabilities from independent tests of significance (1), the value of the combined P becomes .06.

defect, but that it may be due in large part to a wandering of attention or a lessening of motivation engendered by the tedium and boredom of an eight-hour run in confined and uninspiring quarters. This hypothesis gains support from other data collected during the same runs reported in full elsewhere (5). The subjects periodically rated themselves with respect to their feelings of sleepiness, boredom, fatigue, depression, irritability, and general well-being, and with respect to their motivation, attention, etc. Quite generally an end-spurt in the direction of better adjustment occurred in these subjective measures at the last period, when the subjects knew that their ordeal was almost over.

It must not be thought, however, that there are no physiological effects from an eight-hour exposure to the mild anoxia encountered at a

Table 2

Comparison of Mean Syllable Articulation Values for Control and Altitude Runs

Period	I	II	III*	IV
Average Time Elapsed (in Hours)	$3\frac{1}{4}$	$2\frac{1}{4}$	$4\frac{3}{4}$	$6\frac{3}{4}$
Mean for Control Run	31.0%	34.2%	33.7%	34.2%
Mean for Altitude Run	28.7%	27.6%	28.1%	32.7%
Difference (Con.-Alt.)	2.3%	6.6%	5.6%	1.5%
t Value	0.4	1.6	2.0	0.03
P (Probability that Difference is Due to Chance)	.70	.14**	.07**	.77
Con./Alt. $\times 100$	108%	124%	120%	105%

* High protein standardized lunch intervened between periods II and III. It was started after an elapsed time of $3\frac{3}{4}$ hours and was finished within $\frac{1}{2}$ hour.

** When these probabilities are combined by Fisher's technique for combining probabilities from independent tests of significance (1), the value of the combined P becomes .06.

simulated altitude of approximately 10,000 ft. There was a quite general increase in sleepiness, feeling of fatigue, depression, and headache, and a decrease in the general feeling of well-being for the altitude runs as a whole, as compared with the control runs. Furthermore, there was quite reliable and impressive evidence that the angioscotoma, which is relatively free from subjective influence, increased in magnitude during the same altitude runs for the same subjects. These data are also reported in full elsewhere (6). It should likewise be borne in mind that the sensitivity of any criterion of speech intelligibility can be increased by an increase in the initial difficulty of the stimulus material at sea level. It is entirely possible that a clearer indication of diminished capacity to hear speech sounds might have been obtained had the stimulus words been

set at a lower sound level. It will be recalled that the sound level and the stimulus materials employed in this study were the same as those employed in an earlier study which failed to reveal any reliable decrement in speech intelligibility for an exposure of one hour to a simulated altitude of 13,600 ft., though a reliable decrement appeared for a similar exposure at 16,900 ft.

Summary

1. Using the method and materials employed in an earlier study in collaboration with C. P. Seitz (4), twelve subjects were tested for their ability to perceive standard speech sounds at four periods during an eight-hour exposure to the mild anoxia encountered at an altitude of approximately 10,000 ft., simulated in a nitrogen dilution chamber.

2. The decrement in speech intelligibility at altitude was very slight and unreliable at the $\frac{3}{4}$ hr. period; it was nearly reliable at the $2\frac{1}{4}$ hr. and $4\frac{3}{4}$ hr. periods; but there was a marked lessening of the altitude effect at the last period, $6\frac{3}{4}$ hrs. after entering the chamber.

3. The subjects' ability to overcome the mild deterioration in performance exhibited in the middle of the run in an "end-spurt" suggests that the apparent loss of efficiency at the altitude and sound level employed is primarily due to subjective factors such as wandering attention and boredom. The subjects did in fact report that there was a greater increase in sleepiness and boredom generally in the altitude runs than in the control runs. However, evidence derived from another aspect of the study reported elsewhere (6) indicates clearly that there is not a general absence of physiological involvement: there was on the contrary a reliable and progressive enlargement of the angioscotoma during the prolonged altitude exposure. Nevertheless, it seems improbable that significant losses in speech intelligibility will occur on prolonged bombing missions at altitudes of the order investigated; for with properly functioning sound equipment the sound level is much higher than the one employed in this study. It is possible, however, that the subjective factors mentioned may cause errors in speech perception.

Received April 15, 1945.

References

1. Fisher, R. A., *Statistical methods for research workers*. London: Oliver and Boyd, 1936.
2. Fletcher, H., and Steinberg, J. C., *Articulation testing methods*. Bell Telephone Laboratories, Reprint B-436, November 1929, pp. 41, 42, 46.
3. Seitz, C. P., and Smith, G. M., Auditory sensitivity under conditions of anoxia; a study of speech intelligibility, *J. Aeronautical Sciences*, 1942, 9, 478-480.

4. Smith, G. M., and Seitz, C. P., Speech intelligibility under various degrees of anoxia, *J. appl. Psychol.*, 1946, **30**, 182-191.
5. Smith, G. M., The effect of prolonged mild anoxia on attention, irritability, boredom, and other subjective factors. *J. gen. Psychol.*, 1946.
6. Smith, G. M., Seitz, C. P., and Clark, K. B., Variations in the angioscotoma in response to prolonged mild anoxia. *To be published.*
7. Smith, G. M., Clark, K. B., and Hertzman, M., The relation between changes in the angioscotoma and certain Rorschach signs under prolonged mild anoxia. *To be published.*

Studies in International Morse Code: V. The Effect of the "Phonetic Equivalent"

F. S. Keller, I. J. Christo, and W. N. Schoenfeld

Columbia University

The present study originated in two closely related ideas, independently formulated and suggested almost simultaneously by B. F. Skinner and M. Wertheimer (1). Each suggestion arose from the examination of a preliminary outline of a training method described earlier in this series (2), in which the Signal Corps "phonetic equivalents" of the alphabet were employed to identify the individual signals of International Morse code.

With respect to this system of identification, both Skinner and Wertheimer proposed that phonetic equivalents be chosen on the basis of their formal similarity to the signals themselves, without discarding the distinctive cue furnished by the initial letter. Skinner pointed out that the natural tendency of the student to "echo" the auditory signal prior to the explicit written response might be utilized by providing equivalents having sufficient formal overlap with the signals to encourage the arousal, through "verbal summation," of the equivalent itself which, in turn, provides the letter cue. "Thus . . . (S) will no longer lead to the echoic 'di-di-dit', which is of little value, but to 'Sicily' which gives the letter."

Wertheimer's proposal was similar in suggesting the use of "structurally appropriate" equivalents, that is, "words which have the same rhythm, inner grouping, accentuation, length-hierarchy" as the signals, to facilitate the learning, aid against forgetting, help in recall, and avoid confusion of the signals with one another. Wertheimer emphasized the "structural isomorphism" of signal and word, and reported the results of an exploratory experiment which apparently demonstrated the advantage of isomorphic equivalents in both learning and retention.

On the strength of these proposals, it was decided to adapt the general idea to the code teaching procedure referred to above. The tentative list of equivalents offered by Wertheimer for nineteen of the signals; those offered by Skinner for the entire twenty-six; the Signal Corps equivalents; and the words finally selected for the present experiment, are shown in Table 1. The final selection was made as follows: about seventy-five experienced code students were asked to vote on the iso-

Table 1

Words Used as Morse Code Signal "Equivalents"

Note: No equivalents are used for digit signals.

Alpha- bet	Code Signal	Signal Corps	Skinner	Wertheimer	Final Selection
A	..	Able	Around	Ahoy	Around
B	Baker	Beautifully	Boomalacha	Beat Germany
C	Charlie	Chattanooga	Coca Cola	Casa Blanca
D	---	Dog	Dominic	Daintily	Dog did it
E	.	Easy	Eek	Ebb	Eek
F	Fox	Federation	Forestation	Federation
G	---	George	Gold goggles	Gamekeeper	Gamewarden
H	How	Hilly-billy	Helter-skelter	Hilly-billy
I	..	Item	Itchy	—	Itchy
J	----	Jig	Jemima's jam	—	Japan sand man
K	---	King	Kangaroo	—	Kangaroo
L	Love	Legitimate	Los Angeles	Liberia
M	--	Mike	Ma-ma	Mainstay	Ma-ma
N	--	Nan	Naughty	Nasty	Nazi
O	---	Oboe	Oh-oh-oh	Oh my dear	Oh-oh-oh
P	Peter	Prefer posies	Police station	Police station
Q	----	Queen	Quadruplicate	—	Quadruplicate
R	---	Roger	Revoluting	Removal	Revolver
S	...	Sugar	Sicily	Sicily	Sicily
T	-	Tare	Toot	Tea	Toot
U	---	Uncle	Unafraid	Uncle Sam	Unafraid
V	Victor	Victory now	Victory soon	Victory now
W	---	William	Without funds	With all might	Without arms
X	X-ray	Xylophone band	—	Excellent work
Y	----	Yoke	Yankee yacht club	—	Yankee rampart
Z	Zebra	Zoology	—	Zulu did it

morphic suitability of the Skinner-Wertheimer words, together with a number of alternative equivalents. These words, which were presented two or three times in connection with their respective signals, were pronounced with a syllabic emphasis that was calculated to enhance the similarity of word and signal without providing a marked distortion of the former. The signals were transmitted at the same rate as used in actual training. While the equivalents finally selected may not wholly meet Skinner's or Wertheimer's standards, it was felt that they should reduce training time if the basic idea were sound, and that refinement might be postponed until their superiority was demonstrated. The aim of the experiment was not, of course, the confirmation of Wertheimer's results, since the latter were obtained by a different method—one quite impracticable as a general training device.

Nineteen Columbia College undergraduates, ranging in age from seventeen to twenty-three years, were used as subjects in the present experiment. All were inexperienced in code; and all were given fifty minutes of daily training, as a code class, Monday through Friday, throughout the learning period.

In evaluating the influence of the new equivalents, the results of two other experiments are available for comparison. In both of these, undergraduate code classes were used, and the training procedure was identical with that of the present study except that Signal Corps equivalents were employed. The first (3) will hereinafter be called Experiment I; the second (5), Experiment II; and the present study, Experiment III.

The criterion of mastery in Experiments II and III was set as three successive 100-signal runs in each of which the student made no more than five errors (either of substitution or omission). The average number of runs up to the criterion in Experiment II was 23.2 (S.D., ± 12.8); in Experiment III, the average was 22.9 (S.D., ± 8.95). The similarity of results with the two groups of subjects is equally clear when the cumulative progress curves for Experiments II and III are compared (see Figure 1). It is evident that, insofar as speed of learning is concerned, neither set of equivalents has the advantage.

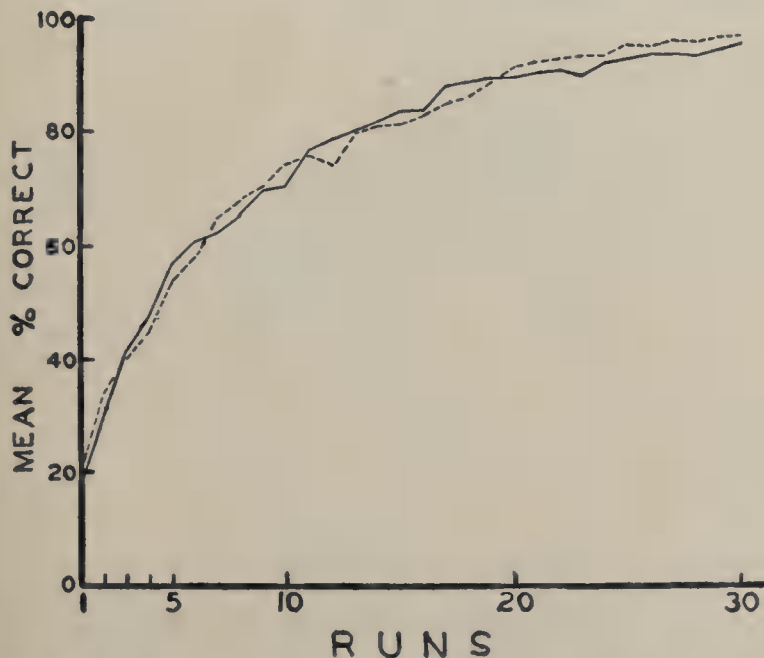


FIG. 1. Progress curves for the students in Experiments II (solid line) and III (broken line). The data plotted represent the average per cent of correct responses for the two groups on successive training "runs" of 100 randomized signals each. Prior to the first run, the 36 signals were identified, hence the curves do not start at zero; 95% correct performance was accepted as mastery. The curve for Experiment III reaches the 95% criterion on run 25, that for Experiment II on run 30.

Table 2

Rank Order Correlations (with PE's) of the 36 Signals in Experiments I, II, and III

Note: The correlations are for rank order of signal difficulty, as based upon substitution, omission, and total errors.

	Exp. I-Exp. II	Exp. I-Exp. III	Exp. II-Exp. III
Substitution Errors	$+.85 \pm .03$	$+.78 \pm .05$	$+.78 \pm .05$
Omission Errors	$+.93 \pm .02$	$+.93 \pm .02$	$+.92 \pm .02$
Total Errors	$+.95 \pm .01$	$+.90 \pm .02$	$+.92 \pm .02$

In Table 2 are presented the rank order correlations of the thirty-six signals, based on substitution, omission, and total errors in Experiments I, II, and III. These correlations indicate that (1) the rank orders for total and omission errors are practically identical in the three experiments, but, (2) in terms of substitution errors, the new equivalents have an effect, as shown by the depressive action of Experiment III upon the correlations. (The stability of the omission errors probably arises from the fact that these errors occur mainly in the very early runs, before the equivalents exercise their influence, and while discriminative failure due to stimulus generalization is primary.)

A better notion of the shift in substitution errors may be obtained from Table 3 wherein are given the rank order correlations between the

Table 3

Eight Cases in which Substitution Errors Common to a Signal in Experiments II and III were Ranked as to Frequency and Correlated

Signal	Rho \pm P.E.
P	$+.42 \pm .19$
F	$+.58 \pm .15$
G	$+.24 \pm .21$
L	$+.16 \pm .19$
U	$+.01 \pm .21$
B	$-.07 \pm .22$
R	$+.03 \pm .22$
4	$+.84 \pm .12$

main substitution errors in Experiments II and III for eight of the thirty-six characters which were selected because the correlations for them could be based on at least ten substitutions, with each substitution occurring at least ten times. Thus, for "P", there were twelve identical substitution errors, each made at least ten times, in the two experiments, giving two rank orders which correlated $+.42 \pm .19$.

Rank order correlations, of course, mask differences in absolute frequencies. For example, it was found that "C" was substituted for "F" twenty-nine times in Experiment II and 136 times in Experiment III. A percentage comparison revealed that this substitution accounted for nine per cent of all the substitutions for "F" in Experiment II, and forty-nine per cent in Experiment III, the difference being highly significant. Other cases of shift in absolute frequency which met the test for significance of percentage differences included the "K" substitution for "U", the "F" for "C", the "C" for "F", and the "U" for "K". A possible reason for these substitutions will be offered below.

It has been argued elsewhere (5) that the difficulty of signals in International Morse code is primarily a matter of stimulus generalization; that certain signals, due to the possession of common properties, give rise to identical responses, thus to "errors" in code reception. The importance of the auditory stimulus similarity is undeniable, but it is now clear that, under the conditions of training that prevailed in these experiments, the response to a signal is in some measure due to another factor—namely, the nature of the word employed to identify a signal. More generally, it may be said that, subsequent to the presentation of a signal-stimulus, there takes place some form of activity on the part of the code learner which works in a supplementary fashion to determine the final, written response.

Both student reports and experimenters' observations support the assertion that, regardless of the specific training procedure, the beginning code student *does something* in the interval between the signal presentation and the written (or spoken) response. He may tap with his pencil, his finger, or his foot; he may whistle softly, or whisper, to himself; he may shake his head and, under strong motivation, seem fairly to bounce in his seat. When no overt activity occurs, he will commonly report sub-vocal, sub-gestural, and visualizing activity, or say that he re-hears or echoes the signal before "responding" to it by writing (or speaking) a character. It is also to be observed that during code mastery the amount of this 'intervening activity' decreases, first for the signals that generalize least, and last for those showing strong and widely distributed generalization.

It has been noted above that, in Experiment III, there was a marked increase in the frequency of certain errors. This may be related to the tendency of the subject to retain his own stress pattern for certain identifying words in spite of the experimenter's attempt to make these words more nearly isomorphic. "Federation" and "Casa Blanca" were frequently confused; so were "Unafraid" and "Kangaroo," and so forth. In effect, these were "poor" isomorphs. It does not, however, appear

likely that the best of isomorphs would entirely avoid the distortion introduced by the already-formed differentiations in the individual subject's language behavior.

In connection with an earlier attempt to analyze the errors made in code learning (4), a few outstanding types of substitution have been described: (1) *reversal* errors, in which a signal is mistaken for its 'mirror image'; (2) *inversion* errors, in which a signal is mistaken for one having the same number of components, but in which dot replaces dash and dash replaces dot; and (3) *dotting* errors, in which a signal is mistaken for one having a smaller or, less frequently, a larger number of dots. A comparison of the frequency of occurrence of these error types in this experiment with the frequency in Experiments I and II reveals a noteworthy decrease in the percentage of reversal errors in this experiment. The attribution of this change to the effect of the new equivalents is supported by the fact that no similar change was observed for the digit signals, which have no equivalents.

The failure of the present study to find any advantage of the isomorphic equivalents with respect to learning time may conceivably be related to the inadequacy of the name-words chosen. Whether better equivalents would bring better results is questionable. The present writers believe that no great advantage is likely to accrue from a more careful choice of words. The speech habits of years' standing—the vocal differentiations already established—in the average student of code are bound to intrude in a fashion that will often lead to erroneous response.

Received June 6, 1945.

References

1. Personal communications to F. S. Keller, November, 1942.
2. Keller, F. S. Studies in International Morse Code: I. A new method of teaching code reception. *J. appl. Psychol.*, 1943, 27, 407-415.
3. Keller, F. S., and Taubman, R. E. Studies in International Morse Code: II. Errors made in code reception. *J. appl. Psychol.*, 1943, 27, 504-509.
4. Keller, F. S., and Schoenfeld, W. N. Studies in International Morse Code: III. The efficiency of the code as related to errors made during learning. *J. appl. Psychol.*, 1944, 28, 254-266.
5. Plotkin, L. Stimulus generalization in Morse Code. *Arch. Psychol.*, 1943, No. 287.

Validity of the Hunt-Minnesota Test for Organic Brain Damage

Rachel F. Malamud

Psychological Laboratories, Norwich State Hospital

Clinical psychologists are frequently called upon to aid the psychiatrist with the practical, clinical problem of determining the presence or absence of organic brain damage in an individual patient. A number of psychologists and psychiatrists have devised tests bearing on this problem, but none have been entirely satisfactory. One of the more recent attempts is the Hunt-Minnesota Test for Organic Brain Damage (1). To what extent this test can fulfill its purpose is the subject of this paper.

The Hunt-Minnesota Test consists of three major divisions: the vocabulary test which is relatively insensitive to brain damage, a group of tests sensitive to deterioration, and a group of interpolated tests. The subject's Stanford-Binet vocabulary score, in relation to his age, determines the score level at which he is expected to perform the more sensitive tests. The deterioration tests, consisting of pairs of words and of designs which the subject is required to associate and later recall or recognize, determines the level at which he is actually functioning. The amount of discrepancy between the subject's expected score and the score he actually makes on the word and design associations is the basis for the diagnosis of brain damage. The discrepancies are indicated by T scores; those T scores which fall higher than a certain critical point are considered to be indicative of organic damage. In an effort, which Hunt (4) describes as "partially successful," to make sure that these high T scores are produced by brain damage alone and not by factors of inattention or poor motivation, he includes in the battery nine short interpolated tests of attention and cooperation. The patient who passes all, or nearly all, of these tests is considered probably capable of being examined validly by the test proper. The final decision on the test's validity for the individual subject, however, is left to the judgment of the examiner.

Hunt's Results

Hunt found his test highly satisfactory in differentiating 33 known organic cases from 41 non-organic control subjects. Using a critical T

score of 68¹ and ignoring one "doubtful" normal, he found that only two organic cases and one non-organic case were misclassified. Even though the total number of subjects used for standardization was small, Hunt (2) felt able to conclude that ". . . the validity of the test battery as a discriminating instrument is statistically established." Subsequent use of the test in the Neuropsychiatric Clinic of the University of Minnesota Hospitals also proved encouraging. Hunt (3) states in a second paper, ". . . T scores above 60 probably justify suspicion of pathology. For practical purposes, a T score of 66 (not 68 as suggested in the manual) should be considered as the 'critical score' dividing the normal from abnormal performances."

Problem

To the members of the Norwich State Hospital Psychological Laboratories the test appeared to have outstanding advantages. It was simple to administer and score; it took only 15 to 30 minutes to administer; it seemed to be satisfactorily validated; and, most important of all, it provided a means for obtaining quantitative evidence of organic brain damage. The test was, therefore, immediately put to use with members of the psychology department as the first subjects. The results from this preliminary use of the test were striking. Six of the ten members of our department were apparently suffering from brain damage. These results naturally led us to question the test's validity. The following problem was formulated for systematic study: *To what extent does the Hunt-Minnesota Test produce "false positives" by designating normals as having organic brain damage?*

Procedure

A total of 64 subjects, all employees of the Norwich State Hospital were tested with either the long or the short form of the Hunt-Minnesota Test. The majority took it as part of a routine battery given to new employees. All subjects fulfilled Hunt's minimum requirements of ability to speak and read the English language, school attendance of at least three grades, adequate muscular coordination and sensory acuity, and a mental age of eight or more. Those given the long form passed all the interpolated tests for attention and motivation. In administering the tests the examiner followed very closely the administration procedure outlined in the test manual and at various times was carefully observed

¹ Hunt actually seems to have ignored the two control cases and one organic case scoring at 68. Of his organic cases, all those falling at or below 67 he considered misclassified, but of the control subjects all those falling at or above 69 he considered misclassified.

not conclude that so large a proportion of normally functioning persons have organic brain damage, neither did we immediately conclude that the test was totally invalid. In the hope of accounting for the discrepancy between our data and Hunt's, we examined our data for possible clues. Two characteristics of our data appeared worthy of consideration.

First, we noted that 29 persons of our total group of 64 had vocabulary scores higher than the upper limit of 32 words for which Hunt says the test is maximally efficient. To determine whether it was these cases which produced the high percentage of "organic" scores, we analyzed only the 35 cases originally within the maximal vocabulary range. Of these we found that 57.1 per cent had "organic" scores. The 29 high vocabulary records studied alone showed 51 per cent, or about the same per cent of abnormality as did all the cases combined. By arbitrarily reducing all vocabulary scores of 33 words or more to 32 words, 42.2 per cent of the total group still remained in the pathological category. Obviously, it was not the superior vocabularies which accounted for the high percentage of "organic" scores.

Secondly, 14 of the 64 cases had been given only the short form. To determine whether these cases unduly affected the total percentage of "organic" scores, we eliminated the 14 records and derived percentages on the remaining 50 "long-form" subjects. Again it was found that the distribution of scores was not greatly changed. Sixty per cent of the "long-form" subjects had "organic" scores. Even by reducing all the superior vocabularies of the "long-form" group to 32 words, 48 per cent still remained in the pathological category. By analyzing the 14 cases given only the short form, and adding to them 23 cases of the "long-form" group on which we had been able to obtain short form scores, we found that 48.6 per cent had "organic" scores. Even after reducing the superior vocabulary scores, 40.8 per cent remained "organic." The short form, then, approximates the long form in the adequacy (or inadequacy) of its discrimination.

In a private communication with the author, Hunt offered the hypothesis that some of our subjects failed the deterioration tests, not because of inability to recognize the proper associations, but because the time limits imposed by the test were too short to allow the recognition to occur. Although we cannot check this hypothesis with our data, it is the present author's impression that it is correct. If the time allowed for recognition had been longer, some of the subjects would probably have improved their scores. If such were the case, however, the test would need revalidation on both normal and organic subjects.

Summary and Conclusions

1. When the Hunt-Minnesota Test for Organic Damage was applied to 64 presumably normal employees of the Norwich State Hospital, 55 per cent had T scores indicating organic pathology.

2. The discrepancy between our results and Hunt's original validation results could not be explained by the fact that our data included cases with very high vocabularies and cases given only the short form of the test.

3. Since the test produces so many "false positives," its validity for diagnosing organic brain damage must be seriously questioned.

Received September 28, 1945.

References

1. Hunt, Howard F. *The Hunt-Minnesota test for organic brain damage*. Minneapolis: The University of Minnesota Press, 1943.
2. Hunt, Howard F. A practical, clinical test for organic brain damage. *J. appl. Psychol.*, 1943, **27**, 375-386.
3. Hunt, Howard F. A note on the clinical use of the Hunt-Minnesota test for organic brain damage. *J. appl. Psychol.*, 1944, **28**, 175-178.
4. Hunt, Howard F. A note on the problem of brain damage in rehabilitation and personnel work. *J. appl. Psychol.*, 1945, **29**, 282-288.

The Hunt-Minnesota Test for Organic Brain Damage in Cases of Functional Depression *

Paul E. Meehl and Mary Jeffery

The University of Minnesota

Among the several tests which have been devised for the detection of intellectual deterioration, one of the most efficient is the Hunt-Minnesota Test for Organic Brain Damage (5, 6, 7, 8).

The author developed this instrument specifically for the diagnosis of organic damage. While the detection and measurement of a decrement in intellectual function, however caused, is an important part of the clinical psychologist's work, methods must eventually be developed for distinguishing between two kinds of deterioration. On the one hand are those which are secondary to "emotional-motivational" factors (e.g. in schizophrenia) and, on the other, those which represent the direct effect of organic central nervous system pathology.¹ Since many varieties of behavior disorders are characterized by a certain amount of psychological deficit, the psychologist will obviously be playing a more significant clinical role if he can make a definite contribution to differential diagnosis (e.g. as between "functional" and "organic" deficit) instead of merely reporting a deviation from the "normal" or "optimal" level.

Such an added report would carry no implication as to the ultimate etiology of the disorder which he has thus labeled as showing either "functional" or "organic" deterioration. Because even if "organic" (endocrine, metabolic, or autonomic) factors should finally be established as primary causes of the development of a schizophrenia, it would still be possible for an observed intellectual deficit to occur as a function of motivation, itself dependent upon the organic factors. In cases with "functional" deterioration, techniques to effectively motivate the patient may cause him to return temporarily to his "true" level, a phenomenon which has often been observed by clinicians. Whereas, in the strictly

* The authors are indebted to Dr. B. C. Schiele and Dr. A. B. Baker, Department of Neuropsychiatry, for their cooperation in this study.

¹ As Hunt has pointed out, test results must be considered as only a part of the evidence required for diagnosis, and must always be interpreted in the light of data from all sources: "Deterioration test scores are thus not a final index . . . but rather a diagnostic and prognostic aid. The extent to which they aid diagnosis and prognosis depends, to a substantial degree, upon the skill and clinical acuity of the interpreting clinician" (8).

"organic" case, no such motivational improvement can make up for a deficit in function directly related to nerve cell destruction, such as occurs in senile dementia or paresis.

In the construction of psychometric devices for detecting organic brain damage, therefore, a difficulty arises because deficit can reflect multiple causation. Consequently, in such psychometric devices, we must either employ tasks which shall yield no decrement for subjects who are merely suffering from test-anxiety, boredom, preoccupation with fantasy, or depressive retardation; or, if this is impossible, we must have means for identifying such special decrements.

The approach suggested by the first alternative is very difficult because possibly it might only be attained by eliminating some degree of test sensitivity to organic deficit. Such a loss of sensitivity might reduce the effectiveness of a test to a point where it would detect only an amount of intellectual loss so gross as to be detectable on other grounds. The ultimate aim of such tests must be to detect *minimal* amounts of damage so that the psychologist can contribute independent evidence of the presence of pathology in the same way as the serologist or roentgenologist can do in cases which are relatively asymptomatic. Psychological tests which merely detect intellectual loss in a person, known on other grounds to be brain damaged, do not contribute maximally to clinical work.

There is reason to believe that this "maximal contribution" is possible because complex intellectual processes are very sensitive to even slight cortical disturbances; and there is already sufficient evidence that the Hunt-Minnesota Test has achieved the increased delicacy desired. Yet just here is the "difficulty" referred to. For, no sooner has the delicacy of the test been stepped up to a point where it can pick up small losses such as those in an early senile change or an undiagnosed encephalitis, than it is also markedly affected by the motivational and emotional factors which are present in other types of cases. In short the increased delicacy is rarely specific for the *kind* of decrement we wish to detect. This would seem to be the crucial problem confronting the clinical psychologist in the field of mental deterioration.

Experience with the Hunt-Minnesota Test at the University of Minnesota Hospitals has demonstrated its high validity as an indicator of organic brain damage. Some of the data have already been published (7), and other validation studies are in progress. Experience, however, in testing cases of depression aroused a doubt that the specificity of the device for organic damage was as great as had originally been hoped. This feeling was first expressed by Hunt, himself, in this journal (8) when he wrote: "In the development of the Hunt test, an attempt was made to provide a special means for identifying those pathologic scores attribut-

able to emotional-motivational disturbances so that the test would then be a specific test for the deterioration associated with brain damage. This attempt has been only partially successful."

Hunt had attempted to identify the scores that he referred to, by including a set of "validity" tests (called *interpolated* tests in his manual) such as digit span, attention, and saying the months forward and backward. His theory was that persons who are disturbed or uncooperative so much as to invalidate their test results would fail the interpolated tests, and thus the examiner would have an index for avoiding an interpretation of deterioration due to organic brain damage. As will be seen from study of the manual, the standard of scoring is extremely lenient; the criterion of invalidity being "failure" on three or more of the nine interpolated tests.

As it stands, the Hunt test showed results that were gratifying with the majority of cases at the University Psychopathic Unit. However, with some persons showing anxiety and depression, high deterioration scores were obtained without other evidence of organic brain damage. Most of these patients seemed quite capable of cooperating as judged by the interpolated tests, usually passing them by a wide margin. To corroborate this clinical impression the present study was undertaken.

It is important that investigations of this sort should avoid unintentionally including subjects already suffering from minimal organic damage. The mere absence of a diagnosis of pathology cannot be taken as proof of normality, without a systematic check in the form of careful history taking and neurological examination. Even using neurology as the criterion, it is unfortunately true that, among the so-called "false positives," an unknown number of persons are actually correctly "positive." However, all one can do is to include only cases which have been neurologically studied and are negative, and then to make the assumption that only a small minority of the group (in the absence of other evidence) have any minimal damage over and above that due to age, for which the Hunt test presumably supplies an adequate correction.

Originally it was intended to obtain retests upon all cases, initially tested during a state of depression, following recovery. This plan was abandoned for several reasons. First of all, research by Arkola (1) indicated the existence of a practice effect of some magnitude. This was apparent even after the lapse of considerably more time than would have passed before "recovery" in our cases. Secondly, the great majority of depressed patients were treated with electroshock therapy which, in itself, may result in unknown amounts of minimal brain damage. The result, then, would have been a combined effect of three variables; two of them (recovery and practice) would tend to lower the T-score by an

indeterminate amount, and the other (shock therapy) would tend possibly to raise it. Accordingly, this scheme of investigation had to be abandoned.

In choice of subjects, several restrictions were necessary. The age limits of 20 to 55 years, for which Hunt claims maximum effectiveness for his test, were imposed. It was required that there be no hint of organic findings or history of shock therapy of any kind during previous episodes. Over a period of eleven months, despite a large number of patients "considered" as subjects, only seventeen subjects fulfilled our requirements. Of these, two were later eliminated because such suggestive signs as slight retinal arteriosclerosis or markedly elevated blood pressure appeared during subsequent neurological study. That the final group of fifteen cases of clearly "functional" depression is small, reflects the extreme care with which cases were selected. The findings, however, are so clear-cut and the Hunt test is being used so extensively that the writers feel further delay in reporting results is ill-advised. Dr. Hunt concurs in this opinion.

The Group

The group studied consists of all in-patient cases with prominent symptoms of depression admitted to the Psychopathic Unit of the University Hospitals from October 1944 through November 1945. Of those who met the required conditions, there were 13 females and 2 males, all between the ages of 34 and 55. The median age was 50 years, with a mean age of 48.7 and S.D. of 6.4 years. Education varied from 7th grade through two years of college, the mean education attained being 10th grade with a S.D. of 2.4 years. Vocabulary level, on the Stanford Binet list, varied from 15 words (M.A. about 13 years) to 29 words (not quite Superior Adult III), with a mean of 22.7 words (Superior Adult I). The *t* of this mean, from a hypothetical supply mean of 20 words, is 2.086 which lies between the 5% and 10% levels of probability.

All of the patients tested had previously received thorough physical and neurological examinations as well as routine laboratory studies. In each case, these were all negative, and no case had a history of head trauma, addiction, or encephalitis.

One patient had a blood pressure of 170/100 but was included because her chart gave three much lower readings for examinations of about 18 months previous. She showed no evidence of cerebral arteriosclerotic changes, neurologically or ophthalmoscopically.

Nine cases were entirely unsedated when tested, and the remaining six were under sedation with either phenobarbital ($1\frac{1}{2}$ grains), sodium amytal (3 grains), nembutal ($1\frac{1}{2}$ grains), or seconal (3 grains). An

unpublished study by Arkola (1) has shown that this amount of sedation with the barbiturates does not produce measurable effects upon Hunt scores, even when administered by injection, and tested at the peak of the sedative effect. Most of the present cases were tested several hours after the oral administration of the sedatives. Furthermore, the mean T-score of the six sedated patients is 66.8 whereas that of the nine unsedated ones is 72.4 (medians 67.5 and 75 respectively). Consequently, it seems safe to assume that these slight degrees of sedation cannot by any means account for the elevations to be reported below.

The staff diagnoses of the fifteen cases were as follows: Involutional melancholia, 5; psychoneurosis, reactive depressive, 4; manic depressive psychosis, depressed, 2; and one each of involutional psychosis, depressed and paranoid; manic depressive psychosis, mixed (agitated); psychoneurosis, mixed (reactive depressive and psychasthenia); and psychoneurosis, anxiety state.

The mean Multiphasic Personality Inventory profile for these 15 cases was as follows: ? 50.8, L 56.5, F 58.8, Hs 66.6, D 86.3, Hy 72.8, Pd 69.7, Pa 72.1, Pt 70.7, Sc 67.4, Ma 52.6, Mf 55.7. In 10 of the cases the depression score (D) was the peak of the profile, and in 11 cases it was above 70. Among the four cases in which D was less than 70, two showed T-scores of 63 on the "lie" scale (L). However, one of these cases was not tested with the Multiphasic until some 55 days after administration of the Hunt, at a time when her psychiatric condition had improved considerably. The median time, elapsing between the administrations of the Hunt and of the Multiphasic was three days, although in two cases an interval of over eight days had elapsed between the administration of the two tests.

It should be pointed out that although all of these patients were depressed in varying amounts, many of them were at some stage of improvement when tested. No patient was tested whose momentary psychiatric condition was such as to preclude his at least claiming ability to cooperate, and apparently doing so. This will be more evident when we later consider the results obtained on the nine interpolated tests.

One case, called "anxiety state" and lacking the word "depression" in her diagnosis, was included because depression, crying, weakness, and insomnia were prominent in her complaints, and because her most marked elevation on the MMPI was on the Depression scale (T-score = 98).

The testing procedure was that described by Hunt in his manual; however, the special urging and explanation required to secure adequate cooperation was possibly more than would be employed routinely. But no actual "coaching" or allowance of leeway in time limits occurred. As

was suggested by the author, the "long form" of the Hunt test was administered. A brief, semi-standardized interview was used following the Hunt test in an attempt to form some impression of the more qualitative aspects of the patient's response to the test situation. The implications of these responses will be discussed below.

The testing was done more or less alternately by the authors, but, due to special circumstances, nine cases were tested by one author and six by the other. Since the mean T-score of these two sets of cases do not differ significantly ($P > .20$), all of the data have been combined for interpretation.

Results

The long-form T-scores of these 15 functionally depressed patients were as follows, in order of magnitude: 88, 87, 87, 87, 83, 75, 74, 73, 69, 68, 65, 62, 55, 44, and 36. The mean of these scores is 70.2 and the median, 73. The sample SD is 15.41 and the best estimate of the supply variability is 15.95 T-score units. Even with a sample this small, it is quite evident that the central tendency of T-scores for depressed patients is considerably above that of the supply mean (of 50) used in interpretation of scores.

Testing the hypothesis that such a sample could have arisen from a population with parameter mean of 50, the Student t is 4.906 which, with 14 d.f., is highly significant ($P < .0002$). We may conclude with confidence, therefore, that the scores of depressed persons cannot be evaluated on the basis of a non-brain-damaged supply mean of 50 T-score.

The obtained estimate of the SD is 15.954, about half again as large as the norm sigma of 10 points. Making use of the fact that the ratio of a sample variance to the supply variance is distributed as χ^2/n , we find a χ^2 of 35.604 which, with 14 d.f., is again highly significant ($P < .008$). It is clear, then, that neither the mean nor the variability of the depressed population can be assumed to be the same as those of the norms.

The confidence belt for the mean (using t) extends down to a T-score of 61.37, using the 5% level of confidence. On the basis of our obtained sample, we may therefore say that the "true" mean of depressed patients is almost certainly not less than about 61, i.e., a full standard deviation above the mean of the general population norms. A similar application of the χ^2 distribution indicates that, at the 5% level of confidence, the "true" SD cannot reasonably be assumed to be less than 12.26 T-score points.

With only 15 cases it was not practicable nor legitimate to make a normal curve fit and test for normality. However, the ω test of Geary (10), employing the ratio of the MD to the SD, was done since it is quite

exact even for this small a sample. The MD of these cases is 12.32, which bears a ratio of .800 to the sample SD. This is almost precisely the mean of the sampling distribution of ω , and there is no reason for assuming that the distribution of scores in the supply is abnormal.

When this approximating assumption has been made, the question arises: How many depressed patients may be expected to show T-scores above the "critical line" of 70? If the sample mean is taken as the best estimate, it is apparent that about half of all cases may be expected to show such spuriously "organic" scores.

Or, more generously, the extreme (most favorable) limits of the confidence belt for the mean and sigma of the supply may be taken. That is, if it is assumed that the true mean is as low as 61.368, as indicated above (a very improbable sampling error), and that the true standard deviation is as small as 12.260, the critical score of 70 is about .704 standard deviations above the mean in such a population distribution. On the assumption of normality, this implies that about 24%, or nearly one in four, depressed persons can be expected to have "pathological" T-scores.

If, as suggested by Hunt in his second article (6), the critical score of 66 were used, the line would be set at .378 sigma above the hypothesized supply mean and therefore 35%, or about one in three, depressed cases would show a "pathological" result. Inspection of the distribution and the mean-median relationship would suggest that, to the extent the assumption of supply normality does not hold, it is because of negative skewness, possibly due to the presence of the rare depressed person whose emotional state leaves his motor and cognitive functions relatively intact. Such a skewness would of course make the proportion of spuriously deteriorated scores even higher.

In summary of these analyses, it is clear that the present sample makes it practically certain that the elevations of T-score in depressed persons cannot be evaluated in terms of the published norms *if* the desired interpretation, that of deterioration due to organic brain damage, is to be made. At the very best, we see that about one in four functionally depressed patients will show scores above the critical line of 70, or about one in three using the score Hunt advises. A much more plausible estimate in terms of the sample statistics is, of course, that about half of the patients will show such elevated scores.

How well do the interpolated tests function in their purpose of detecting such spuriously "pathological" cases? Of the entire group of 15 depressed cases, only one case failed as many as three interpolated tests, Hunt's criterion that the test is invalid. Indeed, only four of the present group failed *any* of the nine interpolated tests; and inspection of the

protocols shows, additionally, that the great majority of the cases were even far removed from the "danger line" on any interpolated test. For those four cases who failed one or more interpolated tests, the T-scores were 87, 87, 83 and 73. The one patient whose test would have been identified as invalid on the basis of interpolated test scores (with a failure of six out of nine) had a T-score of 73.

Arbitrarily, rough weights were assigned to the scores on each interpolated test, and the weights for all nine interpolated tests were summed for each patient. There was no significant relation in our sample between this quantity and the size of T-score ($r = .15$, $P > .50$).

Whether or not the scoring on the interpolated tests could be made more rigorous as a method of solving the present problem cannot be determined from our data. But the good performance of most of the cases and the lack of correlation with T-score, suggests that such a strategem might not work. In order for the majority of functionally depressed patients to fail them, the scoring of the interpolated tests would have to be so rigorous that they probably would begin identifying cases of actual deterioration as "invalidly tested." This seems very likely since these tests have already been used with some success as indicators of deterioration by Babcock (2) and others. However, such a possibility would need to be explored further.

It should be noted that the examiners, on the basis of their previous clinical experience with the Hunt test as used with depressed cases, were able to supplement the interpolated tests in assessing the validity of each test. The test, then, did not "miss" diagnostically as often as the statistics would indicate.

Any reasonably competent clinician would, of course, use his judgment in cases where the psychiatric condition of the patient made invalidity a serious possibility. The examiners, however, would not have been able to distinguish the spuriously high scores adequately here, even though probably influenced by test performance. Before actually scoring the test, each examiner made a rating as to the apparent validity of the testing, trying to exclude estimates of the quantitative results; and to judge, both in terms of the performance as it appeared qualitatively, and in terms of results from the short, post-testing interview. These ratings fell into three categories, namely: probably valid (6 cases), doubtful (4 cases), and probably invalid (5 cases). Dividing the 15 T-scores into three categories from high to low in the same proportions, a chi-square test on the resulting nine-fold table was not significant ($\chi^2 = 5.379$, 4 d.f., $P > .20$).

The results of a short semi-standardized interview, following the administration of the Hunt test, might be discussed briefly. Answers to

the question: "How did you like taking this test?" were rated jointly by the authors, independently of knowledge of test-scores; three categories (favorable, neutral, unfavorable to the test) were used. A chi-square between these ratings and the size of the T-score (9-fold table) was 16.35, which with 4 d.f. is significant at the 1% level. The contingency coefficient based upon this χ^2 is .724, indicating some relationship between how badly the patient performed and his own emotional reaction of disfavor toward the test situation.

An arbitrary weighting of a check list for emotional responses (crying, trembling, etc.) shown by the patient together with subjective judgments by the examiner as to the patient's degree of retardation, motivation, etc., correlated .45 with the T-score which the patient obtained on the Hunt test. That is, high scores were associated with a higher degree of emotional disturbance. With only 15 cases, this correlation lies between the 5% and 10% levels using Fisher's *t*.

Between magnitude of T-score and score on the Minnesota Multiphasic Depression scale, there was an insignificant association ($r = -.13$, $P > .60$).

The four highest T-scores are those of psychotics, but so are the two lowest. On the whole, the diagnoses seem to be scattered randomly among the test scores. The mean score for the nine cases of psychosis was 72.8, and that for the six psychoneurotic cases was 66.3, a difference which is quite insignificant statistically ($P > .40$). Breaking the set of T-scores into "High" and "Low" and then obtaining a chi-square on the resulting fourfold table, again shows an insignificant association between severity of T-score on the Hunt test and diagnosis ($\chi^2 = .028$, 1 d.f., $P > .80$). It should be recalled, however, that the numbers here become so very small that quite possibly the study of larger groups, of psychotics compared with neurotics, would yield a difference.

From these various findings, we may tentatively, with suitable caution because of the small sample, conclude that examiner judgments of validity, amount of upset shown by the patient, diagnosis of psychosis or neurosis, or a measure of depression such as that of the Multiphasic Depression scale, would not enable one to separate valid from invalid testings. It would seem, then, that the best approach is to either avoid giving the test to depressed patients at all, or look upon its results in such cases as indicators of loss in intellectual efficiency without implication of underlying organic pathology.

With a sample this small, such correlations mean little, but it may be worth while to indicate such trends as the relative absence of relation between the T-score and certain other variables. Since the T-score is based upon a deviation from the multiple regression plane (learning score

regressed upon age and vocabulary), one would not expect any relation to exist here. Correlation of T-score with age is insignificant ($r = -.22$, $P > .40$) as is that with vocabulary ($r = -.14$, $P > .50$). The correlation of T-score with maximum grade reached in school is also insignificant ($r = -.30$, $P > .20$).

Qualitative Observations

When questioned, the majority of the patients stated that they felt they could have performed better had they been tested before they became ill. And, it was observed that a number of them showed overt signs of upset such as crying, tremor, and peculiarities of voice and speaking rate. A few expressed a lack of interest in the proceedings, as would be expected in depressed persons. However, all were sufficiently cooperative to be willing apparently to attend to the test material; only two were inclined to admit that they were not really trying very hard.

The explanations patients gave of poor performance varied—that their thoughts tended to be on other things, that they felt too sad to care about the test, and in some cases that they were really trying to make a good showing but simply could not remember adequately. It was not possible, from either the quantitative or the qualitative data at hand, to form any clear hypotheses as to the manner in which depression interferes with the intellectual output.

However, it is likely that the simple fact of retardation could lead to a considerable elevation of T-scores, considering the rather split-second timing which the Hunt test employs. Preoccupation with “other things” is, of course, a possibility; but few would admit to this and, indeed, the examiners’ impression is that this was not a very real factor, considering the more-than-adequate performance of the great majority of the cases when taking the interpolated tests.

Considering the foregoing, the writers are convinced that, on the basis of the subject’s behavior in the testing situation, the examiner cannot adequately judge whether psychiatric upset is seriously impairing validity.

Summary

The Hunt-Minnesota Test for Organic Brain Damage was administered to a group of 15 persons with functional depressions, of whom nine were psychotic and six neurotic. All of these cases were between the ages of 34 and 55 years, and were neurologically and serologically negative for organic brain damage. None of them had a history of alcohol or drug addiction, head trauma, or encephalitis. All were cooperative to the extent of being willing to take the test and to apparently

pay attention to the stimulus materials. Only one of the 15 was disturbed so greatly as to fail as many as three of the interpolated tests, and 11 subjects did not fail any of them. The findings were:

1. The mean T-score of the entire group was 70.2, with a SD of 15.41 points. Both the mean and the standard deviation differ significantly from a hypothetical supply with a mean of 50 and a SD of 10.

2. By the setting up of confidence belts for the estimation of population mean and variance, it is shown that at the very least, one can expect about one in four functionally depressed patients to have "pathological" scores ($T > 70$); or, setting the critical score at 66, about one in three patients.

3. The best estimate is that about half of functionally depressed patients may be expected to show scores over 70 on the Hunt test.

4. It is not possible, from the external manifestations of the patient's emotional disturbance, for the examiner to separate "valid" from "invalid" testings.

5. It is concluded that the Hunt-Minnesota Test for Organic Brain Damage, as it now stands, is not entirely specific for organic brain damage. Significant scores on this test obtained upon cases with depression as an important component of their illness cannot be interpreted except as a decrement in intellectual function of undetermined etiology.

It would be a mistake to extend this interpretation to the test scores of all patients with functional disorders, however, for over half of Hunt's original standardization group was composed of such cases. The mere presence of psychiatric involvement, as in a severe psychoneurosis, is by no means sufficient to invalidate the results, as will be shown by data soon to be published. However, examiners should interpret with caution a significant Hunt score which is obtained on a patient depressed to a considerable degree.

Received February 9, 1946.

References

1. Arkola, A. *The effect of sodium amytal upon performance on the Hunt-Minnesota test for organic brain damage.* Unpublished M.A. Thesis, University of Minnesota, 1945.
2. Babcock, H. *An experiment in the measurement of mental deterioration.* Arch. Psychol., No. 117.
3. Fisher, R. A. *Design of experiments.* London: Oliver and Boyd, 1937.
4. Fisher, R. A. *Statistical methods for research workers* (Eighth edition). London: Oliver and Boyd, 1941.
5. Hunt, Howard F. A practical, clinical test for organic brain damage. *J. appl. Psychol.*, 1943, 27, 375-386.

6. Hunt, Howard F. *The Hunt-Minnesota test for organic brain damage*. Minneapolis: The University of Minnesota Press, 1943.
7. Hunt, Howard F. A note on the clinical use of the Hunt-Minnesota test for organic brain damage. *J. appl. Psychol.*, 1944, **28**, 175-178.
8. Hunt, Howard F. A note on the problem of brain damage in rehabilitation and personnel work. *J. appl. Psychol.*, 1945, **29**, 282-288.
9. Hunt, J. McV. *Personality and the behavior disorders*. New York: Ronald Press, 1944.
10. Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.

Book Reviews

Smith, May. *Handbook of industrial psychology*. New York: Philosophical Library, 1944. Pp. 304. \$5.00.

The preface states that "this little book is not intended to be a detailed chronicle of psychology from the industrial standpoint, but to provide an introduction to the subject for those who are in some way responsible for others, or who have to get on with others."

Neither psychologist nor lay reader would feel after reading this book that his supervision or understanding of workers had been improved. American psychologists will be left wondering whether industrial psychology in England lags far behind that in America or whether the author has inadequately presented work accomplished in her country. Lay readers who note the author's defensiveness in the preface with regard to general acceptance of industrial psychological research, are likely to conclude that the fault lies with psychologists for having devoted the major part of their time to insignificant problems.

The opening chapter gives a historical survey of work preceding modern industrial psychology, and similar material is found throughout the entire book. Much of this material will be new and interesting to many readers, who may wonder, however, at the fragmentary nature of modern work as compared with shrewd observations made several centuries ago. Considerable emphasis is placed on fatigue and environmental conditions such as light, temperature, noise, and hours of work. In this respect the book deviates from the current American trend away from sensory aspects of industrial psychology.

References to American work are conspicuously few. In the field of motion and time studies the work of Ralph Barnes, the Gilbreths, and F. W. Taylor is mentioned. Other references identified by the author or recognized by the reviewer as American are limited to brief mentionings of work by V. V. Anderson, J. Goldmark, Elton Mayo, Munsterberg, Roethlisberger and Dickson, and the National Research Council. Numerous references are for the period during and immediately after World War I and may, consequently, give lay readers the impression that little advance has been made during the past ten or twenty years. The references provide, however, an excellent list of the publications of the Industrial Research Board as well as other British studies; all of which are too little known in America.

Lack of organization is obvious in the author's theories and classi-

fications as well as in the plan of the book as a whole. There is no index, which fact makes it difficult for the reader to gather together all material on a given subject. Many excellent observations and insights appear in the book, but they are scattered and given no emphasis. Interspersed with these are numerous platitudes.

As a whole the book will be disappointing to American psychologists, and the field of industrial psychology will be disappointing to lay readers who judge the field by the book.

Clifford E. Jurgensen

*Minneapolis Gas Light Company,
Minneapolis, Minnesota*

Practical handbook for counselors. Chicago: Science Research Associates, 1945. Pp. 160. \$1.50.

This handbook is directed primarily at counselors in secondary schools.

Handbooks, if encouraged, could easily become substitutes for sound training for counselors. This would be unfortunate because they are of necessity superficial and skimpy. There are many examples of this in the present handbook. On page 54, only three paragraphs are used for the discussion of the Technique of the Interview. In Chapter 5, is "an annotated list" of tests. Not only is the list very limited but also biased as may be seen in the contrasting annotations for the Kuder and the Strong, page 44. Furthermore, in regard to the Kuder the following erroneous statement occurs. "This inventory measures an individual's interests in nine *occupational* areas. . . ." Does it? The reviewer has never seen any evidence to indicate that this is true.

If a person wants a superficial survey of the field of counseling with no intention of practicing counseling, this booklet would be of some use; but the reviewer believes it is unwise to put such a device into the hands of a naive person who intends to do counseling. Such a person might conclude that he could do counseling.

L. E. Drake

University of Wisconsin

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. *The people's choice. How the voter makes up his mind in a presidential campaign.* New York: Duell, Sloan & Pearce, 1944. Pp. 178.

For the most part, this report analyzes the results of repeated questioning of a panel of 600 respondents in Erie County, Ohio, in relation to the 1940 (Roosevelt-Willkie) presidential election. Results and interpretation are stressed in this report while the most important methodological problems are omitted for treatment in separate reports. The

background material includes a description of the county, a summary of political events, and an analysis of the influences operating during the campaign.

The survey itself collected data on voting intentions, expectations, exposure to propaganda, and the usual information on respondents' characteristics—to mention only a few of the more important subjects covered by the interviews. With the use of a panel technique, changes in voting intentions were followed closely and the reasons for these changes were analyzed.

Any attempt to summarize the results or explain the details, within limited space, would do the study an injustice. The interpretations and conclusions cover a broad field all the way from the contention that there is a bandwagon effect to the discovery that "in-laws" agree less than other family members in presidential preference.

The progress made by this study is striking. Its addition to our knowledge of the voter is enough to justify the study beyond a doubt. Psychologists, sociologists, political scientists, and others with an academic interest in political behavior will find the results valuable. Even practical politicians may find a few applications, but they will have to dig them out themselves: the report includes very little interpretation from the standpoint of practical politics.

One reason for this productivity is the use of the panel technique but the reviewer thinks there are other reasons as well: (1) ingenuity in devising hypotheses to be tested, (2) cleverness in developing tests of the hypotheses, and (3) skillful use of breakdowns.

When a study tackles a difficult practical problem, limitations are to be expected; and their presence is not a reflection on the quality of the research. They are listed in this review merely as problems.

1. Many of the most important analyses are based upon sub-groups, and some of these sub-groups are quite small.

2. It is difficult to tell to what extent the findings in a single county would hold up if tested by a more adequate sample. Presumably the trends, relationships, and processes would be fairly uniform in all geographic areas, but even these can be affected by factors which do vary in different areas.

3. People who change political preference definitely within a short period constitute a relatively small proportion of the total. Thus the group that is most productive for the study of changes is not large enough to stand much further subdivision for purposes of analysis.

One omission is relatively unimportant as far as this report is concerned; but it may be significant in relation to other problems. One

reason given for the selection of Erie County was that for forty years preceding 1940 it had reflected national voting trends quite accurately. Naturally many readers will wonder whether Erie County continued to reflect the national trend for the election covered by this study; but the report is silent on this point.

This study has laid the groundwork for similar surveys. An excellent job of ice-breaking has been done. The reviewer hopes that repeating this survey for subsequent elections will be possible.

Alfred C. Welch

*Knox Reeves Advertising, Inc.,
Minneapolis, Minnesota*

Chamberlin, Dean, Chamberlin, Enid, Drought, Neal E. and Scott, William E. *Did They Succeed in College? A Follow-up Study of the Graduates of the Thirty Schools.* New York: Harper and Bros. 1942. Pp. 291. \$2.50.

This investigation concerns the successful adjustments in college of a large number of graduates of the thirty high schools participating in the well-known eight-year study of secondary education. The high schools, it will be remembered, were freed, by agreement with colleges, from requiring that students enroll in the traditionally required college preparatory subjects. This experiment is, therefore, a test of whether students can succeed without the traditional preparatory subjects, having studied the then-called "progressive" subjects. The colleges involved in the study include almost all types and the curricula chosen by the students covered almost the full range available to them. A "comparison" group of control students enrolled in the same colleges were selected and "matched" with respect to comparable scholastic aptitude scores, sex, race, age, religious affiliation, size and type of secondary school, home community, socio-economic status of family, extra-curricular activities in high school, vocational objective and other such factors. These and other data were collected from the colleges' admissions forms and directly from the students themselves. In all, 3583 students—1826 men and 1757 women—were studied, but only 1475 were matched with control students and studied intensively.

Rather extensive analyses are presented of the adjustments of these 1475 experimental and 1475 control students in colleges, those enrolling in 1936 being studied for four years. The graduates of the thirty schools earned grades which were slightly but consistently higher than those of the comparison group (p. 24—). This slight superiority was found in all subjects except foreign languages. In general, the experimental students selected the same types of college subjects for specialization as did the

comparison group. No marked differences between the two groups were found in the number or percentage placed on scholastic probation because of low grades or in the number of scholastic honors received except in the highest level of aptitude where ten per cent more experimental than control students received honors. This latter point deserves further emphasis because of the claim of the progressive educators that the program removed many of the restrictions and hindrances to learning often operating in the cases of very able students. Numerous additional analyses are presented with respect to comparisons of the two groups in scholastic, personal, social and emotional developments and adjustments in the colleges. *In general*, the results are rather consistently favorable to the experimental students.

This reviewer will not criticize certain minor faults of this significant experiment. It is, however, to be regretted that the experimenters did not find it possible to experiment further, or did not see the possibilities of further experimentation, to make their findings more acceptable to those college educators who still doubt that high schools are capable of determining what are satisfactory instructional materials. A smaller experimental and control group could have been selected for testing with the comprehensive achievement examinations developed by the staff of the Eight-Year Study to measure the special and detailed outcomes of the progressive curriculum. Then still other paired groups could have been selected for testing with standardized achievement tests, such as those produced by the Cooperative Test Service and also those of the Iowa Every Pupil Testing Program. These two supplementary experiments would have further tested the relative contributions of progressive curricula, and their opposites, in high schools.

The experiment reported in this book is a classic one, although, for the most part, college faculties and directors of admissions have not rushed to reform their entrance requirements. Sad as it is to report the fact, we must state that only a few cracks have appeared in the college admissions façade. Faculties continue to require prerequisite subject matter for admission as a freshman and to subsequent advanced courses. The whole basis for prescribing such prerequisite subject matter is discredited by such experiments as this one. But little change in practice is observed, and the recent Harvard report on general education might well have been written without knowledge, or concern, for the *experimental* findings of such studies as this one. *Obiter dicta*, not experimentation, continues to be the *modus operandus* for determination of educational policies. We shall need to accumulate current evidence of the college adjustments of veteran-students admitted *without* traditional prerequisites to prove once again that: (1) exposure to, or lack of expos-

ure to, learning opportunities and (2) legislating admissions requirements on the basis of non-experimental evidence are not sound ways of determining who is eligible for and destined to become a successful college student.

E. G. Williamson

University of Minnesota

Wells, F. L. and Ruesch, Jurgen. *Mental Examiner's Handbook*, Revised Edition. New York: Psychological Corporation, 1945. Pp. vii + 211. \$4.50.

This is a revision of the Handbook published in 1942 which has been found very useful in psychiatric examination. It should be stressed that the Handbook is intended primarily for aid in the examination of "psychiatric" patients, that is to say, patients whose behavior deviates from the normal much more than is found in the practice of most readers of the *Journal of Applied Psychology*.

It is divided into a section on so called "clinical" aspects, which merely attempts to list and more or less objectify the usual psychiatric examination of patients; and a section dealing with somewhat more standardized "tests" of mental functioning such as vocabulary, word association, proverb interpretation, the Kent E-G-Y questions, and so forth. The present edition is an improvement over the previous one, especially with respect to the presence of norms. It is particularly useful in the training of medical students just beginning the study of neuropsychiatry, furnishing a more or less objective set of tasks by which they may assess the patient's functioning without actually requiring a thorough study of psychometrics. In some respects the work might be criticized for lending an impression that certain determinations are relatively simple and easy, such as the listing of Murray's "needs" in the same way as one lists the varieties of orientation. Some psychologists might object to the giving of "mental age" values on certain of the tests for similar reasons. On the whole, this Handbook fills an important place between rigorous psychometrics and all that implies, and the almost wholly unstandardized and subjective "mental status" examination of clinical psychiatry.

Paul E. Meehl

University of Minnesota

Sachs, H. *Freud: Master and Friend*. Cambridge: Harvard Univ. Press, 1944, Pp. 195. \$2.50.

This is an enjoyable book. It is light and informative, partly biographical, partly autobiographical. Little insights into the workings of

the group of people surrounding Freud are liberally interspersed, and it furnishes an excellent description of the social atmosphere in which Freud lived and initiated psychoanalysis.

This book is not an important one for personnel directors, but will be useful reading for the majority of those psychologists who are interested in psychoanalysis.

K. W. Oberlin

Western Electric Company

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Psychoanalytic therapy, principles and application.* Franz Alexander, Thomas Morton French, et al. New York: The Ronald Press Co., 1946. Pp. 353. \$5.00.
- Employment tests in industry and business: A selected annotated bibliography.* Hazel C. Benjamin. Princeton: Industrial Relations Section, Princeton University, 1945. Pp. 46. \$.50.
- Music and sound systems in industry.* Barbara Elna Benson. New York: The McGraw-Hill Book Co., Inc., 1946. Pp. 124. \$1.50.
- The successful employee publication.* Paul F. Biklen and Robert D. Breth. New York: The McGraw-Hill Book Co., Inc., 1946. Pp. 179. \$2.00.
- Student personnel work in the postwar college.* Willard W. Blasser, et al. Washington, D. C.: American Council on Education, 1945. Pp. 95. Gratis.
- Manual of child psychology.* Leonard Carmichael. New York: John Wiley & Sons, Inc., 1946. pp. 1496. \$6.00.
- Our teen-age boys and girls.* Lester D. Crow and Alice Crow. New York: The McGraw-Hill Book Co., Inc., 1946. Pp. 365. \$3.00.
- Counseling methods for personnel workers.* Annette Garrett. New York: Family Welfare Association of America, 1945. Pp. 187. \$2.00.
- Cats in a puzzle box.* Edwin R. Guthrie and George P. Horton. New York: Rinehart & Co., Inc., 1946. Pp. 67. \$1.50.
- Twentieth century psychology.* Philip L. Harriman, et al. New York: The Philosophical Library, Inc., 1946. Pp. 710. \$6.00.
- Through a dean's open door.* Herbert E. Hawkes and Anna L. Rose Hawkes. New York: The McGraw-Hill Book Co., Inc., 1945. Pp. 242. \$2.50.
- Human welfare and industrial efficiency.* L. S. Hearnshaw and R. Winterbourn. Wellington, New Zealand: A. H. and A. W. Reed, 1945. Pp. 169. 7s. 6d.
- Adolescence and youth: the process of maturing.* Paul H. Landis. New York: The McGraw-Hill Book Co., Inc., 1945. Pp. 483. \$3.75.
- Stone walls and men.* Robert M. Lindner. New York: The Odyssey Press, Inc., 1946. Pp. 496. \$4.00.

- Principles of dynamic psychiatry.* Jules H. Masserman. Philadelphia and London: W. B. Saunders Co., 1946. Pp. 322. \$4.00.
- The social problems of an industrial civilization.* Elton Mayo. Boston: Division of Research, Harvard Business School, 1945. Pp. xvi + 150. \$2.50.
- The neuroses in war.* Emanuel Miller. New York: The Macmillan Co., 1945. Pp. 250. \$2.50.
- Industrial training and testing.* Howard K. Morgan. New York: The McGraw-Hill Book Co., Inc., 1946. Pp. 225. \$2.50.
- How to keep a sound mind.* Revised edition. John J. B. Morgan. New York: The Macmillan Co., 1946. Pp. 394. \$3.00.
- Men at work.* C. A. Oakley. London: University of London Press, 1945. Pp. 301. 8s. 6d.
- Occupational information.* Carroll L. Shartle. New York: Prentice-Hall, Inc., 1946. Pp. 339. \$3.50.
- Propaganda, communication, and public opinion. A comprehensive reference guide.* Bruce Lannes Smith, Harold D. Lasswell, and Ralph D. Casey. Princeton: Princeton University Press, 1946. Pp. 435. \$5.00.
- Psychiatry in modern warfare.* E. A. Strecker and K. E. Appel. New York: The Macmillan Company, 1945. Pp. 88. \$1.50.
- Thorndike-Century beginning dictionary.* E. L. Thorndike. Chicago: Scott, Foresman and Co., 1945. Pp. 645. \$1.60.
- Interviewing for NORC.* National Opinion Research Center. Colorado: University of Denver, 1945. Pp. 154. \$2.00.
- The Carnegie Foundation for the advancement of teaching.* Fortieth Annual Report. New York: The Carnegie Foundation for the Advancement of Teaching, 1946. Pp. 130. Gratis.

Journal of Applied Psychology

EDITED BY: DONALD G. PATERSON, UNIVERSITY OF MINNESOTA

Consulting Editors

UL S. ACHILLES, *Psychological Corporation*; WALTER V. BINGHAM, *A.G.O., War Department*; ROOLD E. BURTT, *Ohio State University*; ARTHUR I. GATES, *T. C. Columbia University*; HN G. JENKINS, *University of Maryland*; IRVING LORGE, *T. C. Columbia University*; INN MCNEMAR, *Stanford University*; WILLARD C. OLSON, *University of Michigan*; MES P. PORTER, *Swarthmore, Pennsylvania*; EDWARD K. STRONG, JR., *Stanford University*; ORRIS S. VITELES, *University of Pennsylvania*; JOSEPH ZUBIN, *N. Y. Psychiatric Institute*.

Table of Contents

<i>Psychological Corporation's Index of Public Opinion</i> : H. C. LINK	297
<i>Studies in Job Evaluation: IV. Analysis of Another Point Rating Scale for Hourly-Paid Jobs and the Adequacy of an Abbreviated Scale</i> : C. H. LAWSHE, JR., AND S. L. ALESSI	310
<i>Output Rates Among Butter Wrappers: II. Frequency Distributions and an Hypothesis Regarding the "Restriction of Output"</i> : H. F. ROTHE	320
<i>Relation Between Scores on Certain Standard Tests and Supervisory Success in Aircraft Factory</i> : A. Q. SARTAIN	328
<i>Test Validation on Remote Criteria</i> : D. G. HUMM	333
<i>The Development and Standardization of a New Type Test of Peripheral Vision</i> : J. McCLURE	340
<i>Statistical Laboratory for Vision Tests at Purdue University</i> : S. E. WIRT	354
<i>Motor Performance of Normal Young Men Maintained on Restricted Intakes of Vitamin B Complex</i> : J. BROZEK, H. GUETZKOW, O. MICKELSEN, AND A. KEYS	359
<i>Standardization of a Test of Hand Strength</i> : M. B. FISHER AND J. E. BIRREN	380
<i>Time Appreciation Test</i> : J. N. BUCK	388
<i>Relation of Iowa Silent Reading Test Scores to Measures of Aptitude and Achievement</i> : R. W. KILBY	399
<i>The Relationship of College Board Examination Scores and Reading Scores for College Freshmen</i> : H. E. PEIXOTTO	406
<i>Book Reviews</i>	412
<i>New Books, Monographs, and Pamphlets</i>	418

Published Bi-monthly by The American Psychological Association, Inc.

Prince and Lemon Sts., Lancaster, Pa., and

Massachusetts and Nebraska Aves., NW, Washington 16, D. C.

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the Act of March 3, 1879
Copyright, 1946, by The American Psychological Association, Inc.

Journal of Applied Psychology

Vol. 30, No. 4

August, 1946

The Psychological Corporation's Index of Public Opinion

Henry C. Link

The Psychological Corporation, New York City

This survey of attitudes is the fourteenth in a series begun in February 1937. It supplements the Psychological Barometer, a series begun in September 1932 and conducted quarterly with 10,000 personal interviews,—the oldest poll of its kind in existence.

The present study involved 5,000 personal interviews during April 1946 by 479 interviewers in 125 cities and towns, and represents a true cross-section of the urban and small town population. Details as to questionnaires, socio-economic groups, sex, etc. are given at the end.

This fourteenth survey deals with attitudes toward industrial and political issues in an election year as sampled in April 1946.

Do Employers and Unions Have Equal Rights?

The following two questions were asked to throw into bold relief the social responsibility of labor unions on the one hand and of businessmen on the other:

Q. "Do you believe that workers and unions have the right to strike when wages and working conditions don't suit them?"

Q. "Do you believe that businessmen have a right to shut down their factories and stores when labor conditions and profits don't suit them?"

The right of workers to quit their job at any time has long been taken for granted, but in recent years this right has also come to be identified with organized labor unions, as many polls have shown. The right of owners and managers to shut down their plants as a step in dealing with labor unions has been widely questioned. When a large manufacturer recently closed his plants with the statement that he could no longer operate them under conflicting union and government controls, his action was sharply condemned. The answers to the above questions were:

Have Workers and Unions a Right to Strike?

Answers	Total	Socio-Economic Groups			
		A	B	C	D
Yes	63.7%	65%	62%	62%	68%
No	29.3	28	31	30	25
Uncertain	7.0	7	7	8	7
Total interviews	5,000	500	1,500	2,000	1,000

By more than 2 to 1, people answered that workers and unions had the right to strike. While all income groups were about equal in answering "yes" there were substantial numbers who answered "no." Indeed, the largest proportion who said "no" was in the middle income groups. Even among union members, 20% answered "no" though 74% said "yes." When it comes to the right of employers to shut down their factories and stores, the results were quite different.

Have Businessmen a Right to Shut Down?

Answers	Total	Socio-Economic Groups			
		A	B	C	D
Yes	49.5%	60%	57%	46%	41%
No	43.6	32	36	47	53
Uncertain	6.9	8	7	7	6
Total interviews	5,000	500	1,500	2,000	1,000

Evidently the sense of fair play or equality exercises some influence because almost 50% answered "yes." However, this percentage was not nearly so large in the C and D groups as in the A and B groups, whereas in the first question it was about the same in every income group. The results by union and non-union members were:

Businessmen Have Right	Union Members	Non- Union
Yes	43%	52%
No	51	41
Uncertain	6	7

Even among union members, a large proportion concede to employers the right to shut down. The questions raised by these results are: What is the social responsibility of labor unions and businessmen? Do they, as unions or corporations under the law, have unlimited rights to strike? If not, what are the limitations?

The Effect of Changing the Order of Two Questions

Much has been written about the wording of questions but little about their order. The order of the above two questions was reversed in the two forms of the questionnaire. The differences were statistically significant, being four to five times the probable error of 1% for a sample of 2,500 interviews. However, the magnitude of the differences was slight:

	Where Question on Unions Came First	Where Question on Businessmen Came First
Workers have right to strike	65.9%	61.6%
Workers do not have right	27.7	30.9
Businessmen have right	52.3	46.7
Businessmen do not have right	41.3	46.0

67% Favor Stronger Laws to Regulate the Unions

The large majority of the American public favor stronger laws to regulate the unions. If our survey had included the farm population, this majority would probably have been even larger. One aspect of this problem was touched on by the following question:

Q. "If a candidate for Congress promised to support stronger laws to regulate the unions, would you vote for him or against him?"

Answers	Total	Socio-Economic Groups			
		A	B	C	D
Vote for him	66.7%	74%	70%	65%	62%
Vote against him	19.7	16	18	21	22
Uncertain	13.6	10	12	14	16
Total interviews	2,500	250	750	1,000	500

Even more significant than the fact that the large majority favor stronger laws to regulate the unions is the finding that a large proportion of union members are also of this opinion. The answers by union members vs. non-union members were:

Answers	Union Members	Non- Union
Would vote for	60%	69%
Would vote against	26	18
Uncertain	14	13

The CIO—PAC Election Purge

Since the CIO through its Political Action Committee has announced its purpose to purge a large number of political candidates, the following question has become especially timely:

Q. "As you know the CIO unions through their Political Action Committee are trying to elect a lot of Congressmen. If you knew that a candidate for Congress was backed by the CIO, would you be more likely to vote for him or less likely?"

Answers	Total	Socio-Economic Groups			
		A	B	C	D
Less likely	59.4%	67%	67%	59%	45%
More likely	16.0	10	11	14	31
Neither or d.k.	24.6	23	22	27	24
Total interviews	2,500	250	750	1,000	500

The results by CIO, AFL, and non-union members are especially interesting:

Answers	AFL	CIO	Other Unions	Non-Union
Less likely	55%	28%	53%	64%
More likely	20	47	20	12
Neither or d.k.	25	25	27	24

Attitudes toward Certain Types of Candidates

Attitudes toward certain broad classifications of candidates were shown through the answers to the question:

Q. "What kind of man would you be most likely to vote for in Congress: a good business executive; a good labor union leader; a good college professor; a good lawyer; a good politician?"

Answers	Total	Socio-Economic Groups			
		A	B	C	D
Business executive	49.2%	56%	57%	49%	34%
Labor union leader	13.9	4	7	15	28
Politician	12.8	12	13	12	16
Lawyer	11.0	14	10	11	9
College professor	5.2	6	6	5	4
Don't know	7.9	8	7	8	9
Total interviews	2,500	250	750	1,000	500

Businessmen who consider themselves vilified as a class may take some encouragement from these results. Even the union members prefer a good businessman to a good union leader:

Answers	Union Members	All Others
Business executive	34%	55%
Labor union leader	28	9
Politician	14	12
Lawyer	9	12
College professor	6	5
Don't know	9	7
Total interviews	686	1814

Two Groups of Election Issues

Aside from specific questions on certain issues, a list of six possible issues was presented to one-half of our sample and another list to the other half. In respect to each item the person was asked whether he thought it "very important or not so important for Congress" to take action indicated. Then, after expressing himself (or herself) on the six items, he was asked which he considered most important.

Issues	Very Imp't	Not Imp't	D.K.	Most Imp't
Reduce taxes	57%	35%	8%	14%
Give bonus to veterans	73	23	4	30
Pass laws to regulate unions	73	18	9	28
Reduce Government control of business	53	32	15	15
Lend 4 billion to England	29	58	13	6
Lend 1 billion to Russia	22	65	13	1

Loans to England and Russia are considered least important, while a bonus to veterans and regulating the unions are considered most important. Reducing taxes is not a paramount issue. On the other half of the interviews, the results were:

Issues	Very Imp't	Not Imp't	D.K.	Most Imp't
Cut down the number of Government employees	60%	29%	11%	5%
Reduce the Government debt	77	15	8	12
Strengthen the Army, Navy and Air Force	71	23	6	15
Get housing for veterans and others	94	5	1	41
Check Communism	73	19	8	13
Pass laws against race discrimination in employment	53	37	10	10

All these issues except possibly the last were considered very important with the housing issue well in the lead.

Communism in the United States

In this survey we repeated a question which was started as a trend question in 1937.

Q. "Do you believe the United States is on the way to Communism?"

Answers	Feb. 1937	Oct. 1937	Oct. 1939	Oct. 1941	Apr. 1946
Yes	20%	14%	12%	13%	21%
No	64	64	68	75	65
Uncertain	16	22	20	12	14

This shows a definite trend in convictions which is interesting in view of the current emphasis on the dangers of fascism. A similar question on fascism in the 1937 and 1941 studies showed 9% and 8% respectively answering "yes." The above results for April 1946 by union and non-union groups, are especially interesting:

Answers	Union Members	All Others
Yes, headed for Communism	19%	22%
No, not headed for Communism	66	65
Uncertain	15	13
Total interviews	725	1,775

In the other half of the interviews this problem was posed as follows:

Q. "It is being said that Communism is becoming a dangerous thing in the United States. Do you think this is true or not?"

Communism Becoming Dangerous	Total	Union Members	All Others
True	51.2%	55%	50%
Not true	34.1	29	36
Uncertain	14.7	16	14
Total interviews	2,500	686	1,814

It is noteworthy that union members, among whom Communists are reported to be most active, are slightly more fearful of Communism than the rest of the population. The answers by socio-economic groups are also revealing.

Communism Becoming Dangerous	Socio-Economic Groups			
	A	B	C	D
True	43%	51%	55%	50%
Not true	46	39	30	27
Uncertain	11	10	15	23
Total interviews	250	750	1,000	500

The OPA and Government Controls

Formal and informal polls made recently seem to indicate that an overwhelming majority of people favor continuation of the OPA. Undoubtedly the OPA has successfully identified itself in the minds of the people as the one agency fighting against higher prices. Therefore, the people, when asked about the OPA, naturally favor its continuance. The simple question: "Are you in favor of continuing the OPA?" has become almost synonymous with the question: "Do you want prices kept down?" Whether or not the OPA does keep prices down is a matter for debate. When public opinion polls try to reduce a highly complicated and emotional situation to a simple "yes" or "no" question they are on dangerous ground. In the survey reported here, two different questions were asked, each one with one-half of our sample of 5,000 people. One of these questions was:

Q. "Do you think that the powers of Chester Bowles and the OPA should be increased or decreased?"

Answers	Total	Socio-Economic Groups			
		A	B	C	D
Decrease powers	31.6%	42%	35%	30%	24%
Increase powers	25.6	20	24	26	31
Neither	28.6	31	30	30	23
Uncertain	14.2	7	11	14	22
Total interviews	2,500	250	750	1,000	500

This question, while by no means perfect, allows for more than a simple all or nothing answer. It will be seen that a majority, 54%, want the OPA either continued as at present or with increased powers. The 32% who want its powers decreased do not necessarily want the OPA abolished. What is especially interesting about these results is their uniformity by different socio-economic levels. Over 50% in every economic level want the OPA continued with equal or greater powers. The other half of our sample was asked:

Q. "What is doing the most to increase the cost of living: strikes and wage increases; businessmen trying to raise prices; government regulations and restrictions on prices and materials?"

Answers	Total	Socio-Economic Groups			
		A	B	C	D
Strikes and wage increases	43.1%	54%	44%	43%	37%
Businessmen	27.3	23	24	30	30
Government restrictions	26.7	31	28	26	24
Other causes and d.k.	13.2	11	15	12	15
Total interviews	2,500	250	750	1,000	500

These percentages add up to more than 100 because many people gave more than one answer. When these answers are divided by union members and non-union members we have the following results:

Answers	Union Members	Non-Union
Strikes and wage increases	33%	46%
Businessmen	36	24
Government restrictions	27	26
Other causes and d.k.	12	13

Both the union and non-union members, it will be noticed, attribute about the same degree of responsibility for higher prices to government regulations and restrictions. A higher percentage in both groups blame strikes and wage increases and next to these, businessmen trying to raise prices.

Confidence in Government Declines

Possibly the above results help to account for the decline in the people's confidence in the Government's reconversion efforts. Beginning in October 1941 the following question was asked at six month intervals:

Q. "Who do you think can do the best job in straightening things out after the war: the Government in Washington; Business Leaders; Labor Union Leaders; or others?"

(In April 1946 the wording was changed so that the question was as follows: "Now that the war is over who do you think can do the best job of straightening things out at home: the Government in Washington; Business Leaders; Labor Union Leaders; or others?")

Answers	Oct. 1941	Oct. 1943	Oct. 1945	Apr. 1946
Government in Washington	47%	42%	51%	45%
Business leaders	26	28	22	26
Labor union leaders	5	8	9	6
All three together	7	9	12	7
Others or no opinion	17	17	11	16
Total interviews	2,000	2,500	2,500	2,500

While the highest percentage of people still believe in the leadership of the Government, it has dropped from a bare majority to 45%. Confidence in the leadership of union leaders dropped sharply while confidence in the leadership of businessmen had a corresponding rise. In view of these changes it is especially interesting to classify the answers by union and non-union members.

Answers	Union Members	Non- Union
Government in Washington	44%	46%
Business leaders	20	28
Labor union leaders	13	4
All three together	7	9
Others or no opinion	19	15

The union members have more confidence in union leadership than do the non-union members. However, the percentage of union members who have confidence in business leadership is even higher than the percentage of union members who have confidence in union leadership. With regard to government leadership there is little difference either as between union and non-union members or as between different economic groups. In respect to business leadership and labor leadership the confidence varies by different economic groups as follows:

Socio-Economic Groups	Have confidence in:	
	Business Leaders	Union Leaders
A Owner class	36%	2%
B White collar	30	3
C Skilled industrial	25	6
D Semi-skilled	17	13

People Think Themselves Less Prosperous

Six months of costly strikes, sharp wage increases, many price increases, and a large increase in the number of returned veterans have had their effect. In view of the many arguments about wages, etc., it is particularly timely to know whether people *think* they are more prosperous today than they were two years ago. People's opinions about their prosperity were obtained in response to a question which has now been asked seven times since October 1941:

Q. "Is your family more prosperous (or better off) today than two years ago, less prosperous, or the same?"

Answers	Oct. 1941	Oct. 1943	Oct. 1945	Apr. 1946
Better off	38%	29%	32%	26%
The same	47	46	51	48
Worse off	15	23	15	24
Uncertain	—	2	2	2

Throughout the war years the large majority considered themselves more prosperous or at least as prosperous as two years earlier. This belief was fully borne out by statistics of the Department of Labor which showed, for this period, an increase of 35.5% in the hourly wage rates of industrial workers and a 73% increase in the total weekly pay, compared with a 30% increase in the cost of living. However, the April survey revealed a drop of 9% in those who considered themselves as prosperous or more prosperous as compared with the survey made six months earlier. This drop is reflected among socio-economic groups as follows:

Answers	Socio-Economic Groups							
	A		B		C		D	
	'45	'46	'45	'46	'45	'46	'45	'46
Better off	32%	31%	31%	23%	29%	25%	39%	30%
The same	50	49	53	47	53	50	47	43
Worse off	16	18	15	28	15	24	12	24
Uncertain	2	2	1	2	3	1	2	3

Interestingly enough the answers of union members indicated that they were no better off or worse off than the non-union members:

Answers	Union Members	Non- Union
Better off	26%	25%
The same	46	49
Worse off	26	24
Uncertain	2	2

Although the white collar and salaried workers represented by group B seemed to be feeling the pinch the most, the large majority even in this group still considers itself as prosperous or more prosperous than two years ago.

Attitudes toward Five Pressure Groups

Since highly organized pressure groups and lobbies are playing such a large part in the democratic process, the people's attitude toward such groups becomes increasingly significant. It has been claimed that the most powerful groups in Washington today are the Farm lobby and organized labor. However, the former is not known to the public and the latter is represented by two or more distinct organizations. Therefore we asked:

Q. "Which of the following organizations do you think *well* of and which *not so well* of?"

Answers	Well	Not so Well	Doubtful
The U. S. Chamber of Commerce	65%	11%	24%
The AFL	50	31	19
The CIO	26	56	18
Natl. Assn. of Mfrs.	37	17	46
The American Legion	77	15	8

Those who were "doubtful" in some cases had no definite attitude and in others did not know the organization. The contrast between the attitudes toward the AFL and CIO is noteworthy. It becomes even sharper when broken down by CIO and AFL members. The good will of the American Legion is outstanding.

The Probability of Another War

In view of the tremendous interest in peace and measures for a permanent peace, we repeated a question which we asked first in a depth

study in February 1943 (Link, H. C., An experiment in depth interviewing³ on the issue of Internationalism vs. Isolationism. *Pub. Opin. Quart.*, 1943, 6, 267-279). The question was as follows:

Q. "After this war (or, now that the war is over) do you think that we will make a peace settlement that will last, or do you think that we will have another world war in twenty-five years or so?"

Answers	Feb. 1943	Oct. 1944	Apr. 1945	Oct. 1945	Apr. 1946
Will have another war	43%	54%	51%	59%	62%
Will make a lasting peace	47	28	33	28	24
Don't know	10	18	16	13	14

Q. "Who do you think will be our next enemy?"

Answers by Those Who Said There Would be Another War				
Country Named	Oct. 1944	Apr. 1945	Oct. 1945	Apr. 1946
Russia	29%	27%	37%	45%
Germany	9	6	2	2
Japan	5	3	5	1
England	4	4	3	4
China	1	1	1	1
Don't know	6	10	11	9
Total	54	51	59	62

This reflects a steady and sharp increase in the % who expect another war, except for the April 1945 period which reflected the result of the San Francisco Conference. There is a sharp increase in those who believe that the next war will be with Russia. Of the 62% who anticipate another war, about 72% name Russia as the next foe.

Explanation of the Survey

This survey was made during the first three weeks in April with 5,000 personal interviews in 125 cities and towns representing a cross-section of the urban and small town population. Two questionnaires were used, each with one-half the sample, so that some questions were asked of 5,000 people and others of only 2,500. The number of interviews for each question is given in the tables. All interviews were made in the home, but only one in a family. Half were made with women, half with men.

The interviews were distributed by four socio-economic groups referred to in the previous tables as A, B, C, and D. This distribution was

made in accordance with the socio-economic maps in each locality according to which the local supervising psychologist assigned the calls to be made by streets and blocks. The great differences between the thinking of these various socio-economic groups are shown in some of the tables. These differences, incidentally, are also an indication of the thoroughness with which these interviews have been distributed by socio-economic levels.

Received May 23, 1946.

Studies in Job Evaluation: IV. Analysis of Another Point Rating Scale for Hourly-Paid Jobs and the Adequacy of an Abbreviated Scale

C. H. Lawshe, Jr., and Salvatore L. Alessi

Division of Applied Psychology, Purdue University

Previous studies in this series have analyzed the point rating system of job evaluation adapted by Kress¹ for use by the National Electrical Manufacturer's Association as applied both to hourly-paid and salary-paid jobs in industry, and abbreviations of this same rating scale have been examined and compared with the original system. These studies have identified the same or similar two or three factors which function in this rating system, and the abbreviations of the system have yielded results practically identical to those obtained by the complete scale.^{2, 3, 4}

The present study includes a factor analysis of another point rating scale⁵ for hourly-paid jobs in industry and an investigation of an abbreviation of this system. This system differs from the NEMA plan in that each job is rated by means of more and finer categories or degrees on each of the factors, and that the point ratings are translated into "rating factors" by a logarithmic conversion chart for purposes of assigning monetary equivalents. The jobs, then, are paid not by labor grades but by finely graduated "rating factors" which in turn are multiplied by the common labor wage rate in the community to yield the hourly wage rate for any given job.

The primary purpose of the study reported here was to analyze statistically this particular point rating system as used in an industrial plant. An attempt is made also, to relate the basic factors operating in

¹ Kress, A. L., How to rate jobs and men. *Factory Management*, 1939, pp. 60-65.

² Lawshe, C. H., Jr., and Satter, G. A., Studies in job evaluation: I. Factor analyses of point ratings for hourly-paid workers in three industrial plants. *J. appl. Psychol.*, 1944, 28, 189-198.

³ Lawshe, C. H., Jr., Studies in job evaluation: II. The adequacy of abbreviated point ratings for hourly-paid jobs in three industrial plants. *J. appl. Psychol.*, 1945, 29, 177-184.

⁴ Lawshe, C. H., Jr., and Maleski, A. H., Studies in job evaluation: III. An analysis of point ratings for salary-paid jobs in an industrial plant. *J. appl. Psychol.*, 1946, 30, 117-128.

⁵ The names of the authors of the scale and a specific description are withheld at the request of the company concerned.

this system to those item clusters or factors found to be operating in the NEMA system and its modifications previously reported. Finally, the validity of an abbreviated point rating scale based on a few best items was investigated to determine the extent of differences between the two rating scales and their practical significance.

Procedure

Nature of Plan. The plan calls for the rating of industrial occupations on seven so-called elements: General Schooling; Training Period; Manual Skill; Versatility; Job Knowledge; Responsibility; and Working Conditions. Each job is rated on the first six of these items or elements by means of 10 classes or degrees which vary in point values from element to element. Further, the point differentials are not necessarily equal from degree to degree within the same element. Finally, the element "working conditions" is really 3 elements combined. The rater is required to rate each job on three aspects of "working conditions"; namely, "surrounding conditions," "minor hazards," and "major hazards," and each of these is divided into 3 classes or degrees ranging from "normal" thru "poor" to "very poor." Again, the degrees are weighted differently in each of these three working conditions scales. The point rating for "working conditions" is the total of the three sub-ratings.

The point rating of each job is the total of the point ratings on each element. These total point ratings are translated into a "rating factor" by means of a "conversion chart" which consists of two scales and a curve. The horizontal scale is arithmetic and represents the total point rating values and the vertical scale is logarithmic and represents the "rating factor." By locating the point rating of a specific job on the horizontal scale and following this to the point of intersection with the curve, the "rating factor" for that job can be read at the opposite point on the vertical scale. The basic wage rate for the job is then determined by multiplying this "rating factor" or index by common labor wage rate in the community or region.

Source of Data. Point rating data were obtained from an industrial plant having more than 100 different job classifications. These jobs ranged from foremen to plant laborers.

Procedure. Intercorrelations between the point ratings of each of the seven elements and the total point ratings were computed and a correlation matrix was prepared (see Table 1). This matrix was subjected to Thurstone's centroid method of factor analysis while the "rotation of axes" followed Peters and Van Voohris' technique.⁶

⁶ Peters, C. C., and Van Voohris, W. R., *Statistical procedures and their mathematical bases*. New York: McGraw-Hill Book Co., 1940, pp. 248-278.

Table 1

Intercorrelations of Point Ratings of Each of Seven Items and Total Points in a Job Evaluation System in an Industrial Plant

Rating Scale Items	(A) Total Points	(B) General Schooling	(C) Training Period	(D) Manual Skill	(E) Versatility	(F) Job Knowledge	(G) Responsibility
(B) General Schooling	.660						
(C) Training Period	.866	.496					
(D) Manual Skill	.916	.646	.890				
(E) Versatility	.838	.541	.692	.736			
(F) Job Knowledge	.883	.650	.721	.774	.786		
(G) Responsibility	.919	.646	.776	.796	.762	.871	
(H) Working Conditions	-.013	-.269	-.201	-.177	.000	-.149	-.176

The Wherry-Doolittle shrinkage selection method as described by Stead, Shartle, *et al*⁷ was applied to select items for an abbreviated scale.

Factor Analysis Results

Factor Names. The centroid loadings of the factors as they were derived from the analysis are presented in Tables 2 and 3. Factor I was found to be common to all the elements except "working conditions" in comparable magnitude. The loadings for these elements are all within the range of .795 for "manual skill" to .720 for "training period." The

Table 2

Factor Loadings Before and After Rotation

Rating Scale Items	Before Rotation				After Rotation			
	k_1	k_2	k_3	h^2	k_1	k_2	k_3	h^2
(A) Total Points	.951	.249	-.173	.996	.901	.405	.145	.997
(B) General Schooling	.722	-.123	-.268	.608	.780	.000	.000	.608
(C) Training Period	.875	-.120	.317	.881	.720	.028	.601	.880
(D) Manual Skill	.925	-.145	.246	.937	.795	.012	.552	.937
(E) Versatility	.821	.310	-.104	.781	.746	.443	.162	.779
(F) Job Knowledge	.904	.197	.150	.879	.754	.346	.435	.877
(G) Responsibility	.927	.153	.130	.900	.789	.307	.427	.899
(H) Working Conditions	-.198	-.410	-.213	.253	-.045	-.438	-.243	.253

⁷ Stead, W. H., Shartle, C. L., *et al.* *Occupational counseling techniques*. New York: American Book Co., 1940, pp. 245-252.

elements are listed in Table 3 in order of magnitude of loadings. Since each of these heavily loaded elements seems to pertain to some general skill required of the individuals who perform the job successfully the name given to this factor was "Skill Demands (General)."

Factor II has its heaviest loadings in "versatility" (.443) and "working conditions" (−.438). This factor was named "Job Characteristics" since the two elements represent aspects in the job over which the worker has little control and for which he can do little by way of specific training. While the term "versatility" implies something about the individual on the job, actually the scale seems to pertain to the job itself since the emphasis is almost entirely on the routine or repetetive nature of the job.

Table 3

Three Factors Named with Rating Scale Items Arranged in Order of Magnitude of Loadings

Factor	Rating Scale Item	Loading
I. Skill Demands (General)	(D) Manual Skill	.795
	(G) Responsibility	.789
	(B) General Schooling	.780
	(F) Job Knowledge	.754
	(E) Versatility	.746
	(C) Training Period	.720
II. Job Characteristics	(E) Versatility	.443
	(H) Working Conditions	−.438
III. Skill Demands (Specific)	(C) Training Period	.601
	(D) Manual Skill	.552
	(F) Job Knowledge	.435
	(G) Responsibility	.427

That the element "working conditions" is negatively loaded with Factor II is plausible; the element is none the less representative of the factor and its influence is exerted in the negative direction.⁸ There is no question about the element "working conditions" representing an aspect of the job over which the worker has no control.

Factor III has its heaviest loadings in "training period" (.601) and "manual skill" (.552). "Job knowledge" and "responsibility" elements are also loaded appreciably with this third factor. These elements, especially the first two, seem also to represent a skill demand but it is a more specific type of skill requiring a higher degree of specialized training. Therefore, the factor has been named "Skill Demands (Specific)" accordingly.

⁸ Peters, C. C., and Van Voohris, W. R., *op. cit.*, pp. 270–272.

The only elements not contributing to both of the "Skill Demand" factors are "general schooling," "versatility," and "working conditions." "General schooling" is representative of only "Skill Demands (General)" and "versatility" is identified with the "Job Characteristics" factor as well as the "Skill Demands (General)" as discussed above. The "working condition" element is significantly loaded with Factor II, "job characteristics" only. It is logically an isolated element, particularly since it is shown in the correlation matrix to be negatively correlated with the

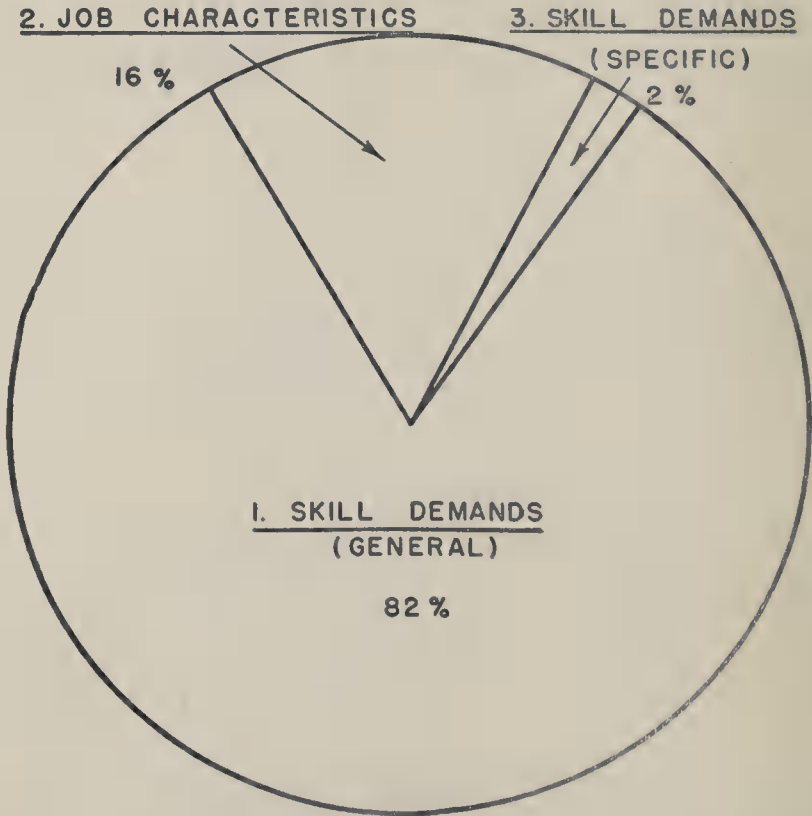


FIG. 1. The relative proportion which each of the factors contributes to the total point ratings of hourly paid jobs in an industrial plant.

total point ratings and the other elements (see Table 1), and because the remaining elements contribute to both of the "Skill Demands" factors.

Factor Significance. Assuming the rotation of axes (Table 2) to be the best possible, the relative proportions that each factor contributes to the total variability may be estimated by squaring and adding the three loadings for the total point rating. Since the communality (h^2) of the three factors for the total point rating element is .996, practically all the variability in the total point rating is shown to be accounted for.

The proportions are represented graphically in Figure 1. Factor I

"Skill Demands (General)" accounts for 82% of the total variability, Factor II "Job Characteristics" 16%, and Factor III "Skill Demands (Specific)" 2%. These proportions indicate the relative extent to which the various factors contribute to the total variability of the total point rating and, in consequence, to the determination of the wage structure, since it is based on the total point rating for each job.

Similarity Between Systems. Although the point rating system analyzed in this study differed from the NEMA system analyzed in previous studies of this series, the basic factors found to be functioning in both systems seem to be comparable. The "Skill Demands" factor found in all previous studies to account for most of the variability in the total point rating has its counterpart in this system in the "Skill Demands (General)" factor which accounts for most of the variability in the total point ratings. The "Job Characteristics" factor ranked second in contributing to the total point rating in both systems again. And, finally a third factor which seems to embrace any specific skill demands, visual, supervisory, or specialized training, accounted for the remaining 2 or 3% variability in the total point ratings of both the NEMA system and of this system.

The Abbreviated Rated Scale

Elements Selected. The Wherry-Doolittle selection method was applied and three elements were selected for an abbreviated scale. The three elements are "responsibility," "manual skill," and "working conditions." These three elements properly weighted correlated with the criterion total point rating .983. If the Wherry-Doolittle process had been carried further to include a fourth element, the multiple correlation coefficient would have increased by only .005.

The one element which correlates highest with "total points" is "responsibility," the coefficient being .919 (see Table 4). When the "manual skill" element is added the multiple correlation coefficient is

Table 4

Correlation Coefficients Between Ratings on Selected Scale Items and Total Point Ratings with Standard Errors of Estimate

Selected Items	Correlation Coefficients	Standard Errors of Estimate (Total Points)	Percentage
Responsibility	.919	52.3	38.6
Responsibility plus Manual Skill	.968	33.3	25.1
Responsibility plus Manual Skill plus Working Conditions	.983	25.0	18.8

increased to .968; and when "working conditions" is added the multiple correlation coefficient increases to .983.

Accuracy of Prediction. Table 4 includes the standard errors of estimate for predicting the total point rating from either one, two or three items in an abbreviated scale. If only one item were used to predict total point ratings, the estimates for approximately two-thirds of the

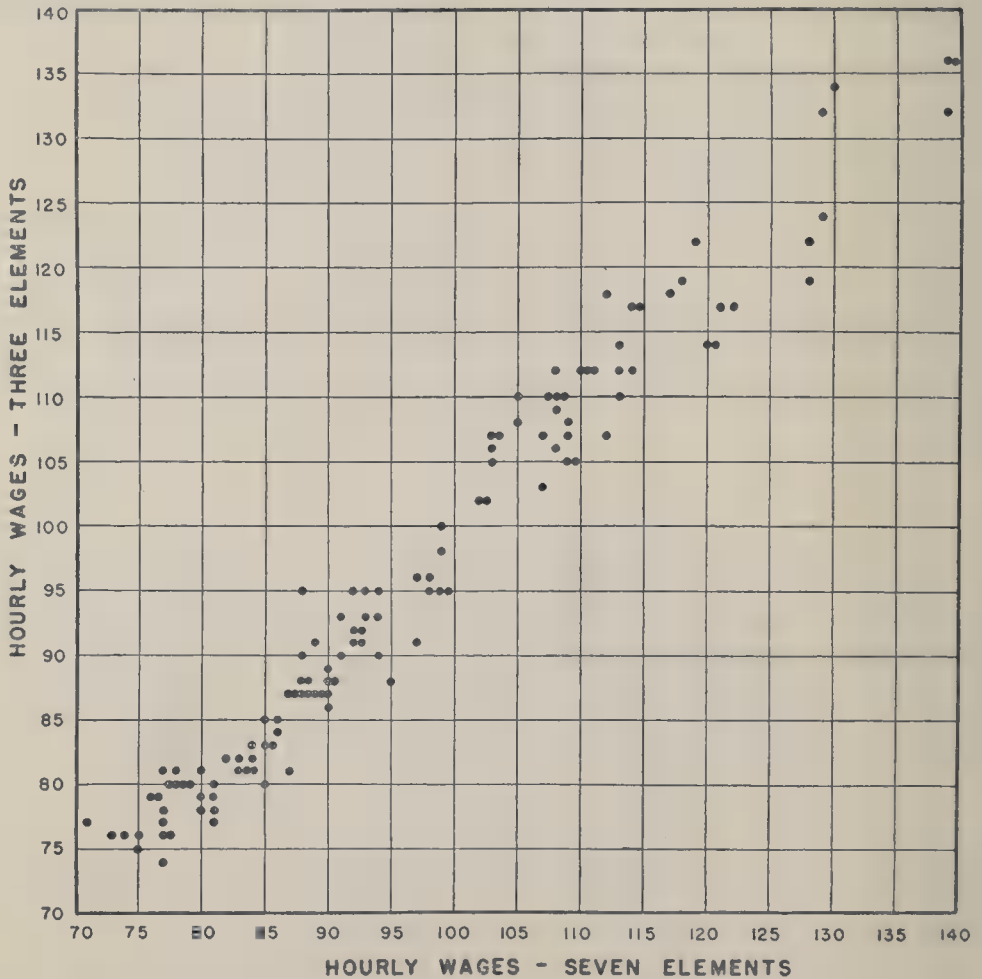


FIG. 2. Scattergram showing relationship between hourly rates based upon total points (seven items) and rates computed from three items.

jobs would be within 52.3 points of the total point ratings based on all 7 elements. If the best three items were used, the estimated total point ratings for approximately two-thirds of the jobs would be within 25.0 points of the total point ratings based on all seven elements. The percentage figures in Table 4 for each of the selected items indicate the proportional size of the errors in terms of the standard deviations of the total point distribution.

Coefficient of Multiple Determination. The coefficient of multiple determination is .964. Hence, it can be assumed that the three elements, "responsibility," "manual skill" and "working conditions," in an abbreviated rating scale contribute or account for 96.4% of the total variation in total points while the other elements account for the remaining 3.6%.

Application of Abbreviated Scale. In order to test the practical application of the three element abbreviated scale, the changes that would occur if only the three elements were used were analyzed and the extent to which comparable results would be obtained was examined. For this purpose the regression equation for predicting "total points" from the "responsibility," "manual skill" and "working conditions" was used. The equation is:

$$X_{TP} = 43.0 + 1.7 X_R + 2.7 X_{MS} + .8 X_{WC}$$

Point ratings on "responsibility," and "manual skill" and "working conditions" were substituted in the regression equation formula for each of the jobs in the plant to obtain the computed ratings. These computed ratings and the total point rating for all seven elements of the original scale were translated into "rating factor" indices and these were multiplied by the going hourly-paid wage rate (\$.71) for common labor in that vicinity to yield hourly-paid rates for each job. The money values (in cents) predicted by the abbreviated scales were then plotted against money values derived from the application of the original seven element rating scale in the scattergram in Figure 2. The grid lines in the scattergram indicate five cent intervals in hourly-paid rates.

Wage Differences. Table 5 summarizes the differences in "predicted" and "actual" wage rates paid for the jobs in the plant according to the

Table 5

Differences in Wage Rates as Computed with All Seven Elements and the Three Selected Elements

Difference in Cents	No. of Jobs	Cumulative Frequencies	Cumulative Percentages
7	3	122	100
6	5	119	97
5	6	114	94
4	13	108	89
3	19	95	78
2	30	76	62
1	32	46	38
0	14	14	11
Mean Difference = 2.3 cents.			

abbreviated and original rating systems. The greatest difference between "predicted" and "actual" wage rates is seven cents, and only 3 jobs showed this large a deviation in wage rates as determined under the two systems. Ninety-four per cent of the jobs were paid within a five cents differential and the average difference is only 2.3 cents.

Since about half of the time a wage rate paid by the abbreviated rating scale would "miss" the "actual" wage rate paid by less than three cents, it seems safe to assume that for practical purposes the abbreviated scale could be substituted for the original. This is significant when it is realized that a certain rate may be paid at one or another time to employees on jobs ranging thru several "rating factor" indices, because of seniority in the plant, length of service on the job, or for other reasons that company policy may deem desirable or reasonable. The relative unreliability of point rating scales further minimizes the significance of the slight differences in results yielded by the abbreviated and original scales.⁹ *In terms of practical operation, the two systems seem almost identical.*

Summary and Conclusions

The operation of a point rating system of job evaluation in an industrial plant was examined statistically by means of Thurstone's centroid method of factor analysis. By means of the Wherry-Doolittle technique an abbreviation of the scale was set up and compared with the original scale. The following findings are supported:

1. There are three primary factors operating and they jointly account for practically all (96%) of the variability in total point rating.
2. The factor which contributes most to the variance in "total points" is the "Skill Demands (General)" (82%). This factor represents a general skill requirement for the individual who can do the job successfully.
3. The second factor which contributes 16% to the total variance in "total points" is "Job Characteristics." This was so named because it includes certain aspects of the job with which the worker must contend and over which he has little control.
4. The third factor is "Skill Demands (Specific)" and accounts for 2% of the variation in "total points." This factor represents the skill demands of the job of a more specialized nature.
5. These results support the finding of previous studies in that the basic factors found operating in the NEMA system were also revealed in this plan.

⁹ Lawshe, C. H., Jr., *op. cit.*

6. The abbreviated scale was made up of the "responsibility," "manual skill" and "working conditions" elements. Ratings on selected elements have a multiple *R* of .983 with the total point ratings.

7. If the abbreviated scale were used in this plant, only three of the jobs would be displaced as much as seven cents, 94% of the jobs would not be displaced more than five cents, and about half the jobs would not be displaced more than 3 cents in the wage structure of the plant.

8. The practical significance of the slight differences are further minimized by the flexibility of the wage system and the probable unreliability of the ratings.

9. The abbreviated scale would yield results practically identical to those obtained with the original scale and would greatly reduce the time required as well as the complexity of the rating process.

Received April 20, 1946.

Output Rates among Butter Wrappers: II. Frequency Distributions and an Hypothesis Regarding the "Restriction of Output" *

Harold F. Rothe

Stevenson, Jordan and Harrison, Inc., Chicago, Illinois

In an earlier paper on the analysis of output rates among butter wrappers (9), some of the problems involved in the use of work curves were discussed. The methodology of this study and the conditions under which the data were obtained were described in detail. The purpose of the present paper is to describe a further analysis of those data and to relate them to other industrial problems.

There appears to be a constantly increasing appreciation of the magnitude of the range of individual differences in abilities and in output rates by industrial managers. Several writers have described distributions of the output rates of industrial workers. For the most part these are bell-shaped distributions; many of them appear to be skewed, and no one appears to have tested these distributions for normality. According to Evans (3), the ratio between the outputs of the greatest and the least producers or the "best" and the "worst" workers on any job runs about 3 or 4 to 1 for those operations where the pace is set by the workers rather than by the machines. That is, the "best" worker produces 3 or 4 times as much as does the "worst" worker in a given period of time. To this statement by Evans should be added the further qualification,—when the operators are approximately equally experienced. Ratios of this kind for light manual operations have been published by Hull for heel trimmers, 1.4 to 1, and for bottom scourers, 2 to 1 (7, 35). Hull also reported a ratio of 5 to 1 for spoon polishers, citing a report by Farmer and Brooke. Inspection of the original data, however, reveals that this distribution included both "experienced" and "semi-experienced" operators, and that when the semi-experienced group was given one week's training the ratio dropped to 2 to 1 (4). Tiffin presented polygons for electrical fixture assemblers and persons burning, twisting, and

* This study was made in partial fulfillment of the requirements for the degree Ph.D. in psychology at the University of Minnesota. The writer wishes to thank his co-chairmen, Professors Donald G. Paterson and Miles A. Tinker, for their many helpful suggestions. He is also grateful to Mr. John Brandt, President, and the employees of the Minneapolis plant, Land O'Lakes Creameries, Inc., whose cooperation made possible this investigation.

soldering the ends of insulated wire. The ratios of these groups appear to be approximately 2.5 to 1 (12, 4 ff.). Stead and Shartle have published output histograms with ratios of 4 to 1 for card-punch operators, 1.5 to 1 for experienced lamp-shade sewers, and 3 to 1 for inexperienced lamp-shade sewers (11, 75 ff.).

A few writers have discussed frequency distributions in connection with "restriction of output." Ford hypothecated that a negatively skewed distribution was "a fairly certain index of organized restriction" (5, 78 ff.). Yoder later presented a similar hypothesis that the distribution of output data would be skewed negatively and have a lower average and less variation if production were restricted than if it were unrestricted (14, 282 ff.). Neither of these writers presented evidence to support their hypotheses. Bliss found the distribution of earnings for a group of bench workers to be positively skewed, and he attributed this skewness to a lack of motivation (2).

Bedford attacked this problem in a different manner, making one histogram for each individual shoe factory worker studied (1). These distributions, based on factory records over a 20-24 week period, showed a tendency toward a common modal rate of production for both fast and slow producers, with the fast ones being positively skewed and the slow ones negatively skewed. He also found less variation about their own means for the fast individuals. He proposed using these two phenomena, common mode and differential skewness for both groups, as an objective measure of "restriction of output."

The Present Problems

Primary interest in the present investigation was in the work curves and their stability. It appeared worthwhile, however, to analyze the obtained data in terms of their frequency distributions, to add these distributions to the too-small number of such published distributions, and to attempt to relate these data to practical industrial problems. It was hoped to test the hypothesis of Ford and Yoder, and the contradictory explanation of Bliss, with the data. The smallness of the present sample of operators (eight women) made this impossible.

The problems that were investigated may be listed briefly in this manner: (1) to obtain individual and group frequency distributions of output rates for industrial operators; (2) to test these distributions for normality; and (3) to investigate any possible phenomena about these distributions in relation to a measurement of "restriction of output."

"Restriction of Output" Redefined

It is well at this point to indicate more clearly what is meant by the term "restriction of output." This term is generally used to indicate

that workers are producing at a rate lower than the rate they are capable of maintaining over a long period of time without suffering any ill effects. Ford wrote in 1931 that restriction was probably present in over 90% of the major industries of the United States (5, 78 ff.), and Mathewson wrote that restriction among organized workers is believed so common as to be almost universal and therefore not requiring detailed inquiry to ascertain its presence (8, 4 ff.).

The present writer believes the term is ill-advised because of its one-sided emotional connotation. That is, the term as it is now commonly used suggests that the reasons for this phenomenon lie wholly within the activities of the workers and outside of the actions of management. Experience in this war has shown that a large number of industrial phenomena are functions of both management and labor, and that nothing is to be gained by blaming, even by suggestion, one or the other of the two parties.

From a psychological point of view, when workers restrict their output they feel that they have more to gain by producing below their optimum than they have from producing at their optimum. The problem is basically one of *incentives*. If the incentive is great enough, the workers will work at an optimum rate, that is, producing as much as they can, steadily, over a long period of time, without endangering their health or decreasing their away-from-the-plant activities. If the incentive is not great enough they will work below this optimum, that is, they will "restrict" their output.

The problem of incentives is a problem for both management and labor, and also for industrial scientists, to solve. If the term "restriction of output" is replaced by the term "ineffectiveness of incentives," joint action by all parties may be achieved more easily. That more desirable term is used from this point on in the present paper.

Results

Frequency distributions were made for each operator showing the various rates of output they made at any time during the two week period, with the exception of the first day of the study.¹ The data for this day were higher than for any other day and were therefore excluded from all analyses as reflecting a spurious phenomenon. These distributions of output data are presented in Table 1.

¹ In Part I of this report the only data used covered five days. The data used here cover the two-week period. Alternating job assignments for the operators forbade the construction of usable work curves covering more than 5 days, but all the data were used in these distributions regardless of alternating duties.

Table 1

Frequency of Occurrence of Output Rates for each Operator, in terms of Pounds of Butter Wrapped in Fifteen-minute Periods, Grouped into Class Intervals of Three Pounds

Class Interval	Operator Number								Total
	1	2	3	4	5	6	7	8	
65-67	—	—	—	—	—	—	—	—	—
62-64	1	—	—	—	—	—	—	—	1
59-61	4	—	—	—	—	—	—	—	4
56-58	23	—	—	—	—	—	—	—	23
53-55	47	—	—	—	—	8	—	—	55
50-52	47	—	—	—	—	7	—	—	54
47-49	30	—	—	—	—	32	—	—	62
44-46	14	—	—	—	—	43	—	—	57
41-45	5	1	5	—	2	53	—	—	66
38-40	—	19	8	13	7	15	16	4	82
35-37	—	41	34	31	11	8	27	17	169
32-34	—	62	68	41	32	4	33	36	276
29-31	—	48	54	31	37	—	49	51	270
26-28	—	20	19	13	43	—	36	30	161
23-25	—	9	6	4	15	—	19	13	66
20-22	—	2	—	4	5	—	5	—	15
17-19	—	—	—	—	—	—	—	—	—
Totals	171	202	194	137	152	170	184	151	1361

All of the distributions in Table 1 were tested for normality according to the Chi Squared test, except the data for the whole group which were tested with class intervals of 5 rather than of 3 as shown here. At the 1% level of significance, none of the individual distributions differed significantly from normal. This was also true when the eight Chi Squares and their respective degrees of freedom were added and the test was in terms of a normal deviate with a standard error of 1. Using the method of beta coefficients to test for skewness and kurtosis, the distributions for Operators 1, 2, and 7 were significantly leptokurtic. This apparent discrepancy between the Chi Square and the beta coefficient tests for normality may be attributed to the use of the 1% level in the Chi Squared test,—a procedure that tended to make it difficult for the distributions *not* to be normal.

The distribution of output rates for the whole group, shown in Table 1, tested as significantly skewed in the positive direction. It should be noted clearly, however, that although this distribution contains hundreds of readings, it really refers to a distribution of eight cases only. The influence of Operators 1 and 6 account for the non-normality of this distribution. That is, two operators produced at a sufficiently faster

rate than did the other six operators so that the distribution of group data was positively skewed.

In a manner similar to the above, distributions were also made for each operator and for the group on another one of the operations they performed during the two week period. This operation, briefly, consisted of wrapping one-pound blocks of butter in a manner essentially similar to the manner of wrapping quarter-pound blocks. In that operation, the distribution for all operators tested as normal according to the Chi Squared and the beta coefficient methods, except the distribution of Operator 4, which was non-normal at the 1% level on the Chi Squared test.²

In summary, it appears proper to say that these distributions are bell-shaped and tend to approximate the normal frequency distribution. This conclusion must be limited to these data and a more general statement cannot be made at this time.

In connection with Bedford's hypothesis of measuring "restriction" it is noteworthy that although six of the eight operators have fairly close means, there is no common mode for both fast and slow producers, nor are the distributions for the fast workers skewed positively and those for the slow workers skewed negatively.

The ranges of inter-individual differences and intra-individual differences were obtained for both wrapping operations that have been described. These data are summarized in Tables 2 and 3.

Table 2

Data on Inter- and Intra-individual Differences in Operation of Wrapping Four Quarter-pound Blocks of Butter in Fifteen-minute Periods

	Operator Number							
	1	2	3	4	5	6	7	8
Mean rate.....	51.5	32.2	32.4	32.3	29.9	44.0	30.8	30.5
Fastest rate.....	64.	41.	43.	40.	41.	52.	56.	40.
Slowest rate.....	36.	21.	22.	20.	20.	32.	21.	23.
Ratio*.....	1.78	1.95	2.41	2.00	2.05	1.63	2.67	1.74

* Ratio of each operator's fastest to her slowest rate.

The ratios of fastest to slowest rate were used as measures of the ranges of intra-individual differences. The ratios between the mean rates of the fastest operator and the mean rates of the slowest operator

² In testing the distributions for this second operation, corrections were made for the small number of cases. Yates' correction for continuity was used in the Chi Squared test (6, 102), and kurtosis was tested by obtaining the standard errors of the *g* coefficients (6, 29).

were used as measures of the range of inter-individual differences in this situation. These latter ratios were 1.69 and 1.77, respectively, for the data in Tables 2 and 3. These latter two ratios were surprisingly alike, considering the small sample used in this investigation, and led to the adoption of the single ratio 1.73 to 1 as expressing the range of inter-individual differences here.

Table 3

Data on Inter- and Intra-individual Differences in Operation of Wrapping One-pound Blocks of Butter in Fifteen-minute Periods

	Operator Number							
	1	2	3	4	5	6	7	8
Mean rate.....	85.8	62.4	58.8	50.8	48.5	72.6	61.2	58.1
Fastest rate.....	109.	78.	80.	92.	76.	91.	88.	92.
Slowest rate.....	68.	48.	30.	32.	32.	52.	34.	37.
Ratio*.....	1.60	1.63	2.67	2.88	2.38	1.75	2.59	2.49

* Ratio of each operator's fastest to her slowest rate.

From Tables 2 and 3 it is seen that only one ratio of intra-individual differences in Tables 2 and two ratios in Table 3 were lower than the ratio of inter-individual differences.

A few other analyses were made of these data and the results of these are summarized below.³ Analyzing the distribution of production rates by days for each operator revealed that: (1) the six slower operators tended to have a common mode for any one day; (2) all operators tended to have a common mode for themselves from day to day (note the straight-line work curves previously described); (3) all operators tended to have a common mean for themselves from day to day; (4) all operators tended to have a constant relative variation from day to day; and (5) the faster operators showed relatively less variation in production rates from day to day than did the slower operators. All of these analyses, of course, were based on a small amount of data and can only be interpreted as suggesting trends.

An Hypothesis regarding the Measurement of the Effectiveness of Incentives

To the extent that the present limited data permitted, it was desired to investigate the plausibility of the various hypotheses concerning the

³ These analyses were made by inspection of many elaborate tables. The complete report of these is contained in the writer's thesis filed at the University of Minnesota Library.

measurement of "restriction of output" or, as it has been re-defined here, of the "effectiveness or ineffectiveness of the incentives" offered for working at an optimal rate.

To make such an investigation required the assumption that the incentives for this particular group of employees were not very effective. The writings of Ford and Mathewson tend to justify that assumption even before any evidence is sought in this or any other situation. One inherent bit of evidence is the fact that these employees were paid a straight time wage, plus overtime, and that no incentive system was in operation. Most books dealing with motion and time studies describe many situations in which output, even among very experienced workers, increases greatly when an incentive system is substituted for a straight time wage. The low ratio between the mean rate of the fastest and the slowest operators also suggests that the incentives were not wholly effective. This ratio is 1.73 in this situation. Wyatt, Frost, and Stock showed that this ratio increases as the incentives to work become more effective (12). Seashore described how the range of individual differences increases when the different persons use different work methods (10). The present employees did use different methods here, and still the ratio was only 1.73 to 1.

This ratio, although large in terms of economic significance, is small when compared to the ratio that is customarily found in laboratory experiments on motor and manual skills. Possibly the larger ratio found in laboratory experiments, often greater than 3 to 1, is a reflection of the strong incentives operative in the laboratory situation. There is an enormous difference in motivation between the student who performs a laboratory manipulation for five or ten minutes and the industrial worker who performs his tasks hour after hour, day after day, and year after year.

There is another respect, too, in which these industrial tasks and ratios may be compared with laboratory experiments, although the data are not very clear-cut in either of these instances. The reliability of most light manual and motor tests is generally 0.90 or higher. This means that any given individual will tend to get very close to the same result on a re-test. But the workers studied here did not show a close correspondence on their re-tests. That is, each individual worker here showed a large range of output rates, so large that, in most instances and for the same operation, the ratio of the range of intra-individual differences was higher than the ratio of the range of inter-individual differences. This would most likely not be true of a typical laboratory experiment on a similar type of manipulation. *This leads to the hypothesis that the incentives to work may be considered ineffective when the ratio of the range of intra-individual differences is greater than the ratio of the range of inter-individual differences.*

This hypothesis of industrial motivation is presented here as a tentative one deserving further investigation. It is derived from a very small number of cases and there are some exceptions to it in these very data. It is derived with the aid of the *assumption* that the incentives were not wholly effective in this situation. There was, however, no clear-cut evidence of "restriction" among these operators. More than that, the extremely higher productive rates of Operators 1 and 6 over the other operators might better be taken as evidence that there was certainly no "organized restriction." But the assumption that the motivation was not very high for most operators here is justified partly by the other writers mentioned above, partly by the impressions recorded by the investigators during the study, partly by the tendency towards a common mean rate of production among six of the operators, and partly by the tendency for the group members to vary their output together over any one day as reported in the previous paper. Further investigation is needed to test the value of this hypothesis.

Received June 25, 1945.

References

1. Bedford, T. The ideal work curve. *J. Industr. Hyg.*, 1922, 4, 235-245.
2. Bliss, E. F., Jr. Earnings of machine tenders and of bench workers. *Person. J.*, 1931, 10, 102-107.
3. Evans, W. D. Individual productivity differences. *Month. Labor Rev.*, 1940, 50, 338-341.
4. Farmer, M., and Brooks, R. W. *Motion study in metal polishing*. London: Ind. Hlth. Res. Bd., Rep. No. 15, 1921.
5. Ford, A. *A scientific approach to labor problems*. New York: McGraw-Hill, 1931.
6. Goulden, C. N. *Methods of statistical analysis*. New York: John Wiley and Sons, 1937.
7. Hull, C. L. *Aptitude testing*. New York: World Book Co., 1928.
8. Mathewson, S. D. *Restriction of output among unorganized workers*. New York: The Viking Press, 1931.
9. Rothe, H. F. Output rates among butter wrappers: I. *J. appl. Psychol.*, 1946, 30, 199-211.
10. Seashore, R. H. Work methods: an often neglected factor underlying individual differences. *Psychol. Rev.*, 1939, 46, 123-141.
11. Stead, W. H., and Shartle, C. L. *Occupational counseling techniques*. New York: American Book Co., 1940.
12. Tiffin, J. *Industrial psychology*. New York: Prentice-Hall, 1942.
13. Wyatt, S., Frost, L., and Stock, F. G. L. *Incentives in repetitive work*. London: Ind. Hlth. Res. Bd., Rep. No. 69, 1934.
14. Yoder, D. *Personnel management and industrial relations*. New York: Prentice-Hall, 1942.

Relation Between Scores on Certain Standard Tests and Supervisory Success in an Aircraft Factory *

A. Q. Sartain

Southern Methodist University

The question of how to select supervisory personnel is frequently one of the most important faced by a business enterprise. Since success as a worker is no guarantee of success in supervision, it is natural that psychological tests should be considered as possible instruments for selection of suitable persons for supervisory responsibilities.

Statement of the Problem

The problem of this study was to determine the extent to which success in supervision in an aircraft factory was predicted by the following standard tests: Otis Self-Administering Test of Mental Ability (Higher Examination); Tiffin and Lawshe Adaptability Test (Form A); Revised Minnesota Paper Form Board; Bennett Test of Mechanical Comprehension (Form AA); Remmers and File How Supervise? (Experimental Edition, Form A); Bernreuter Personality Inventory; and Kuder Preference Record.

Subjects and Conditions of the Experiment

The tests listed above were given to 40 members of supervision in the factory. Thirty-seven of these men were assistant foremen, and three were foremen. Each man was rated by the foreman and general foreman over him (except in the case of the foremen, where it was necessary to secure a second rating by the general foreman, the second rating being obtained about three weeks after the first). Each man was rated on two different rating forms, and the combination of the four ratings constituted the criterion of success.

The Criterion

In setting up the criterion, the ratings on each rating form were converted to standard deviation scores, and the sum of these scores became

* The writer wishes to express his appreciation to the Texas Division of North American Aviation, Inc., for supporting and making possible this study. Special acknowledgment is made of the help of Mr. Ross A. Peterson, Director of Education.

the criterion. An attempt was made in preliminary studies to insure both the reliability and the validity of each rating form. One of these forms (called Form A henceforth) consisted of the seven qualities which had been found to correlate most highly with success as a supervisor, each quality being listed on a separate sheet. In the preliminary study, the correlation between the average of two ratings and the average of two scores or grades given for success on the job was .88 (.84 when new ratings were secured five weeks later), and the correlation between two ratings for each man was .64. The number of employees involved was 43. Thus, it is concluded that Form A was sufficiently reliable and valid to comprise a part of the criterion.

The second rating form (Form B) consisted of ten qualities, all on a single sheet. In the preliminary study ($N = 54$), the correlation between the average of two ratings and the average of two scores or grades on supervisory success was .92. The two ratings correlated with each

Table 1
Correlations between Ratings Constituting Criterion

Ratings	<i>r</i>
Average Rating on A vs. Average on B79
First Rating on A vs. First on B77
Second Rating on A vs. Second on B62
First Rating on A vs. Second on B54
Second Rating on A vs. First on B48

other to the extent of .63. Thus, it appears that Form B was also reasonably reliable and valid.

It should be emphasized that the results just cited were from earlier studies of the rating forms. In the present study the results were hardly so favorable. Table 1 presents the relevant findings for this study. While these correlations are not as high as earlier studies might lead one to expect, they appear to be high enough to indicate that the combined ratings might well serve as the criterion.

Results of the Study

As Table 2 brings out, correlations between the test scores and the criterion were low, in every case so low as to lack statistical significance. (According to Fisher, for a coefficient of correlation to be significant at the 5% level of confidence under the conditions of this study it would have to be .304; at the 1% level of confidence it would have to be .393.¹)

¹ Guilford, J. P., *Psychometric methods*. New York: McGraw-Hill Book Co., Inc., 1936, p. 549.

Table 2
Coefficients of Correlation between Test Scores and Criterion

Test	<i>r</i>
Otis Self-Administering	.04
Adaptability	-.07
Minn. Paper Form Board	.10
Bennett Mechanical Comprehension	-.15
How Supervise?	-.18
Bernreuter Personality Inventory	
B1-N	-.11
B4-D	.12
F1-C	.01
F2-S	.07
Kuder Preference Record*	
Mechanical	.004
Social Service	-.06
Clerical	.003

* The plant was closed before this study was concluded, and the data on the other interest scales of the Kuder test inadvertently destroyed.

These low correlations may be due to a faulty criterion. It seems more probable, however, that the tests simply fail to correlate with supervisory success in this plant.

Correlations were obtained between some of the test scores, and are presented in Table 3. The correlation between the two general mental ability tests (.86) and those between the mechanical ability tests and the general mental ability tests (.33 to .41), as well as that between the two

Table 3
Coefficients of Correlation between Certain Test Scores

Tests	<i>r</i>
Adaptability vs. Otis	.86
" vs. How Supervise?	-.44
" vs. Form Board	.33
" vs. Bennett	.41
Otis vs. Form Board	.39
" vs. Bennett	.37
Bennett vs. Form Board	.31
How Supervise? vs. Kuder Persuasive	.00
" vs. Kuder Social Service	.17
Kuder Mechanical vs. Form Board	.13
" vs. Bennett	.15
Kuder Scientific vs. From Board	.19
" vs. Bennett	.15

mechanical ability tests (.31), are not far different from those found in most similar studies.² The correlation between Adaptability and How Supervise? indicates that general mental ability goes with favorable supervisory attitudes (low scores on this test indicating a favorable attitude) to a moderate degree. Other coefficients are too small to have significance.

Additional Studies

Two other studies were made of the success of the Otis and Bernreuter in selecting supervisors. In one of these, the sum of the scores on both the rating scales was again used as the criterion of success, two ratings

Table 4

Relation of Bernreuter and Otis Scores to Rated Success in Supervision

Test	<i>r</i>
Otis	.16
Bernreuter	
B1-N	-.12
B4-D	.04
F1-C	-.09
F2-S	-.02

Table 5

Comparison of Bernreuter and Otis Scores of Groups of Good and Poor Supervisors

Test or Scale	Poor Group				Good Group				Critical Ratio
	No.	Mean	S.D.	S.D. _M	No.	Mean	S.D.	S.D. _M	
B1-N	29	-127.1	61.20	11.56	24	-146.4	46.30	9.66	1.29
B4-D	29	85.6	48.50	9.16	24	108.3	65.10	13.58	1.39
F1-C	29	-95.0	73.05	13.77	24	-109.3	51.90	10.80	.82
F2-S	29	-36.5	44.20	8.35	24	-38.8	48.90	10.18	.17
Otis	28	101.1	13.03	2.51	24	105.1	10.08	2.10	1.22

on each form being secured on 85 men. Table 4 is based on this study. It is clear that the coefficients are most likely due to chance.

In the second study, 53 members of supervision who were known well to three individuals in management positions were divided into two groups, good supervisors ($N = 29$) and poor supervisors ($N = 24$). The members of each group were selected because there was agreement among those classifying them that they belonged in one or the other group.

² Greene, E. B., *Measurements of human behavior*. New York: Odyssey Press, 1940, p. 257; p. 361.

When the Bernreuter and Otis scores of these two groups were compared, the results shown in Table 5 were obtained. It will be noted that the differences all favor the good supervisors, that is, that they appear to be more intelligent, more stable, more dominant, more self-confident, and more sociable, but that no difference even approaches statistical significance.

Summary and Conclusions

The following tests were administered to forty members of supervision in an aircraft factory: Otis Self-Administering Test of Mental Ability (Higher Examination); Tiffin and Lawshe Adaptability Test (Form A); Revised Minnesota Paper Form Board; Bennett Test of Mechanical Comprehension (Form AA); Remmers and File How Supervise? Test (Experimental Edition, Form A); Bernreuter Personality Inventory; and Kuder Preference Record. Two ratings on each of two rating forms were then secured for each man, the rating forms previously having been checked for reliability and validity, and the sum of the four ratings (reduced to standard deviation scores) became the criterion of success. Test scores were then correlated against the criterion. In every instance the coefficients obtained were too low to be considered significant, the highest one being only .18. It was concluded, therefore, that these tests had little or no predictive value for success in supervision in this plant.

Two additional minor studies corroborating this conclusion in part are also reported.

Received August 31, 1945.

Test Validation on Remote Criteria

Doncaster G. Humm

Personnel Service, Los Angeles, California

The demonstration of the usefulness of psychological tests in an applied situation, such as selection for employment, is a problem of test validation on remote criteria. Test validation on immediate criteria, of course, involves the demonstration of the effectiveness of a test's ability to measure that which it is meant to measure. Thus, a skill test is validated on an immediate criterion if it is shown to be capable of measuring skill; an intelligence test, to measure intelligence; an interest inventory, to measure interest; and so forth.

However, there are some situations in which some demonstration of a test's usefulness to the situation itself needs to be shown. Before an intelligence test properly may be considered a part of the battery for the selection of employees, there must be some demonstration that intelligence is a factor in success on the job and that the test being considered measures intelligence in such a situation. The result is that test validation on remote criteria sometimes becomes necessary.

A rigorous solution of such a problem requires careful consideration of basic assumptions plus the provision for adequate safeguards against the intrusion of extraneous factors. It is, in fact, a complicated scientific experiment which must be set up and carried through with due regard to the requirements of scientific method.

Two major attacks on such an experiment are possible: (1) the employment of logical analysis, and (2) the employment of quantitative methods. Either of these attacks will be successful if it is essentially factual, carefully controlled, and adequately thorough.

The procedure employing logical analysis may well start with a factual consideration of all of the characteristics needed for success. Thus, if the problem is the selection of combat flyers, it will be necessary to isolate the causes of failure and also the factors leading to success. In many cases, the latter factors present opposite phases of the former; but this cannot be taken for granted. Some of the factors leading to success are separate and distinct from those leading to failure. The demonstration of both the qualities required for success and those predisposing to failure must, of course, be proved beyond reasonable doubt. Thus, if intelligence is to be a factor in such a situation, there must be

some demonstration that certain ranges of intelligence are accompanied by success and certain are accompanied by failure. It is, however, sufficient to demonstrate by factual job analysis that the characteristics are important in selection.

The quantitative consideration of such a problem may be attacked in either of two ways: (1) all of the critical factors contributing to or hindering success in the situation except the factor to be measured by the test under examination may be equated and held constant; or, (2) all of the factors may be measured, their interrelationships with success and with each other determined, and the effect of all factors except that under consideration partialled out.

As can be readily seen, it is almost impossible to separate the two general types of attack first mentioned, since logical analysis and quantitative considerations are both considered in the second and since, also, quantitative methods may often be successfully used in the first.

The point to be emphasized, however, is that the task is not simple. The direct correlation between an intelligence test and success on the job, for example, is not justified; since it is possible that other factors may vary to the extent that zero or negative correlation between intelligence and success on the job may be obtained when actually intelligence may be a decisive factor. This is quite likely to happen in jobs which require low-average intelligence, since it has been demonstrated that intelligence too high for the job is equally as handicapping as intelligence too low for the job. As a consequence, the relationship between success on the job in such job brackets and scores in intelligence tests is likely to be curvilinear, with greater tendency to report high intelligence in negative fashion than low intelligence. Incidentally, the same result is likely to happen with regard to skill tests and aptitude tests.

Any attempt to oversimplify the attack on the validation of tests on remote criteria is very likely to bring discredit on the test unjustly.

Let us consider an example of the correct way to examine the selective value of a test, taking, for instance, the Minnesota Spatial Relations Test for assembly workers. If we assume that intelligence, interest, temperament, physical fitness, previous experience, and quality of supervision are all factors of success or failure on this job, we may proceed in one of two ways. In the first, one should provide that these other factors are all so well equated as to permit us to examine the effect of the Minnesota test uninfluenced by their variations. As an alternative, we may set up measures for all of these factors, determine the relationship between these various factors and success on the job and each other, and partial out their effect on the Minnesota test.

Note that the first thing we had to do was to make an assumption as

to the factors important to success or failure on the job. There is, however, no justification for making those assumptions, unless they have been proved by some previous examination. As a consequence, since the second of the two alternatives provided justification for such assumptions, it is probably more likely to return valid results than the first.

One needs only to consider the end results of such a careful consideration of many factors and contrast them with the original correlation between success on the job and the test under consideration to see how dangerous it is to attempt to use the latter measure alone and without examining these other factors.

There is one major drawback to the employment of this second type of attack on the problem. It is that it involves the use of multiple regression equations of high order which must be solved by determinants so complicated that solution borders on the impossible.

Let us use as an example the problem of studying the effectiveness of the Kuder Preference Record in the selection of salesmen. Again, let us assume that intelligence, skill, temperament, physical fitness, previous experience, and quality of supervision are additional factors. There are nine factors to be considered in the components reported by the Kuder, seven additional factors if we use the Humm-Wadsworth Temperament Scale for temperament, and at least one factor for intelligence, skill, physical fitness, previous experience, and quality of supervision. In all, this makes 21 factors to be considered with relation to the criterion of success on the job; a task requiring a determinant of the 21st order, or, in other words, a set of 21 simultaneous equations containing 21 unknowns. While the solution is possible, it is extremely difficult, especially if the error in rounding off is adequately considered.¹

For ready solution, therefore, the problem must be considerably simplified. It is possible to accomplish this by considering the Kuder as a unit on a scale of relative pertinence and the measure of temperament on relative acceptability of the behavior tendencies reported. Some of the other factors then may be eliminated by equating them. For example, only those salesmen may be considered who are physically fit, satisfactory in experience, and well adjusted to their supervisor. In this way, the number of unknowns may be reduced to four and the solution of the determinant simplified.

The point is, however, that these factors have all received consideration. Failure to do this will result in inaccurate findings.

¹ The error in rounding off may lead to very inaccurate results if not adequately considered. Thus, if $50 \pm .5$ is multiplied by $50 \pm .5$, the product equals $2,500.25 \pm 50$. Similarly, if $1,440 \pm .5$ be divided by $.012 \pm .0005$, the quotient is approximately $120,000 \pm 5,426$ or somewhere between 125,426 and 114,574.

Some factors of success and failure are difficult to deal with quantitatively. One of these is the quality of supervision. This is an extremely important and yet a frequently neglected consideration. It is not easily dealt with for the reason that the matter of incompatibility often enters the picture. Thus, a foreman may actually be an excellent foreman and the worker an excellent workman and both may get along well with others on the job, but their two temperaments may clash to such an extent that they seem unable to work together without striking sparks.

Another factor that often enters the picture is that of compensatory mental reactions or the influence exerted on one characteristic by other characteristics. This may be manifested either positively or negatively. Thus, an individual worker may have fairly low skill and yet perform very adequately by reason of the fact that he is painstakingly careful or is highly intelligent or is an assiduous worker. Contrariwise, an individual may have very superior skill; but, because he is too intelligent for the job, may be so bored as not to put forth the effort to manifest that skill.

Compensations in the field of interest are especially frequent. An examination of the interest patterns of any group of workers may indicate, if interest alone be considered, that a number of these workers are misplaced. This may or may not be true. The fact is that many of these individuals may be securing sufficient compensation for their lack of interest in the job by some avocational pursuit which gives adequate expression for that interest, and other incentives for staying on the job may be strong enough to make them wish to remain. Thus, if interest examinations are measured against success on the job, the real part that interest plays in that success is difficult to measure.

One of the drawbacks to the validation of tests on remote criteria is the fact that some factors in success and failure cannot be readily taken into account. Thus, it is very difficult to consider the effect of home conditions, outside social adjustments, worry about finances, and the like. At some times these are very real factors in success or failure of the individual.

On the whole, then, we see that the problem of considering the effectiveness of a test in its applied situation is extremely complicated and difficult. Unfortunately, the psychological literature reports many examples of failure to consider these complications. In many instances, the requirements of scientific experiment are seriously overlooked. If these examples were the exemplification of the work of tyros, a paper such as this perhaps would not be necessary; but the fact is that some psychologists of high rank have used remote criteria without adequate precaution. These examples will illustrate:

1. A report made on the Guilford-Martin Inventory (on the basis of 48 cases) assumed that behavior was an exclusive resultant of temperament (behavior tendency) and neglected the trigger effects of intelligence, skill, and supervision.
2. A study made of the Humm-Wadsworth Temperament Scale (59 cases) reported it as ineffective because, when used alone, it did not differentiate between good and poor salesmen. No controls whatever were set up for experience, intelligence, skill, etc.
3. A critique of the Bernreuter Inventory (on the basis of 95 normal and 329 abnormal soliders) reported that "raw scores were not significantly differentiating." No interrelationships were used, and no other factors, aside from the inventory, were considered.
4. On the basis of two cases and eight months follow-up of one case, it was concluded that the "Rorschach seems to have high validity in this type of employee selection." No other evaluation techniques were mentioned.
5. Seven sub-tests selected from the Bennett, McQuarrie, and O'Rourke were found to have a multiple correlation of $+ .47$ with instructor's ratings of success in mechanical training course of 147 high-school students. No equating of intelligence or consideration of other factors was reported.

We have already pointed out that the problem of test validation on remote criteria will require some logical analysis as well as quantitative treatment. This follows from the fact that applied mathematics requires both a sound knowledge of mathematics and a thorough familiarity with the subject matter. Indeed, a mathematical treatment is no more than a logical manipulation which proceeds from certain assumptions to conclusions. It follows that mathematical treatment must be founded on that good sense which is common to the subject matter. This again is a requirement which often has failed to be met in psychological literature. There are too many examples of psychologists conducting research on the validity of tests without beforehand making themselves thoroughly acquainted with the work which has already been done on the test. Some recent publications on the Bell Inventory, the Kuder Preference Record, the Bernreuter, and other tests will illustrate this point.

It is also to be regretted that we have so many reports that are based on an insufficient number of cases to establish statistical momentum. In fact, all, except possibly one, of the five examples previously quoted illustrate this error.

A suggested remedy for this situation is a more extended use of logical analysis. The writer would be one of the last to decry the use of quantitative method, and yet one of the first to criticize its unwarranted or

superficial use. Mathematics is probably the queen of the sciences;² yet, mathematics does not constitute the only way to think.

In a great many situations, actually, it is a waste of effort to validate a psychological test on such a remote criterion as success on the job, since there are certain qualities which may safely be assumed as being prerequisite for the job. Among such qualities that are at least to a considerable extent desirable are commensurate skill, pertinent interests, good mental health, freedom from anti-social tendencies, and adequate physical health.

A critical analysis of many of the reports on the usefulness of tests will reveal that the validity study might better have been made on an immediate rather than a remote criterion. For example, if we can assume that fine dexterity is necessary for the assembly of watches and we have adequate proof that the Purdue Peg Board is a good measure of fine dexterity, then it must follow that the Purdue Peg Board is likely to prove a valuable member of the test battery for watch assemblers. Similarly, if we can assume that a cashier must be trustworthy and honest and we have a test which validly measures honesty and trustworthiness, then it follows that such a test is very likely to help in the selection of cashiers.

Such a direct attack on the problem is often likely to be more successful than an attack through the use of remote criteria.

In many instances in the industrial situation, the case-study method may be used to an advantage. This is especially true in the study of failures. The writer assisted in such a study. It included 330 problem employees in a public service company. There the case-study method revealed that approximately 6 per cent of the failures were explained on the basis of unfitting intelligence, 6 per cent on the basis of skill, 6 per cent on the basis of physical fitness, 80 per cent on the basis of temperament, and 2 per cent for miscellaneous reasons. Similar case studies of problem employees are very likely to reveal characteristics that are important, at least for their elimination in selection procedures.

If such studies are followed by case studies of outstanding employees, the observations already made may be verified and information of additional importance revealed.

Summary

The component tests of a test battery cannot be directly validated on success in the situation or the job unless the requirements of scientific experiment are rigorously met. This implies either that all other factors except the one under consideration are equated and kept constant or that

² As Eric Temple Bell has said.

adequate mathematical safeguards (including provision for taking care of the error of rounding off) are set up to measure all important variations. Such a task requires both logical analysis and mathematical treatment. Neither may be neglected.

Where the situation or the job is adequately analyzed and where the characteristics needed for the job are clearly established, it is better to validate tests on immediate criteria. That is to say, it is better to ascertain whether or not the tests are effective in measuring that which they are meant to measure. While this requires no less rigor in mathematical treatment, it is such a simpler task that the results are more likely to be found in accordance with the facts.

Received October 2, 1945.

The Development and Standardization of a New Type Test of Peripheral Vision *

John Allan McClure

Division of Education and Applied Psychology, Purdue University

Recent studies in industry have proven the value of proper visual skills in relation to job performance, quality of workmanship and accident experience. Results indicate that one of the visual skills, peripheral vision, not previously included in the research, might prove to be an important factor in jobs requiring this skill, such as crane operator, truck driver, and industrial tractor operator. In order that research may be conducted to determine the importance of peripheral vision in industry the new type perimeter described in this report was developed.

The present study is concerned with the determination of the reliability of the instrument, the relationship of peripheral vision to other visual skills, and the accumulation of standard norms.

Development of the Perimeter

Knowledge of the field of vision and a realization of its limits have been indicated in the literature from the time of the Greeks and Romans. Thomas Young (10) in 1801 made the first accurate study of the field of vision. He listed the outer limits of the normal field at 90° on each side. Purkinje (6) reported in 1825 an outer limit of 100° and 110° with the pupils dilated. Von Graefe (9), in 1855, was the first to report a study of the visual fields for diagnostic purposes. He used a simple campimeter consisting of a small blackboard with a piece of chalk on the end of a wire as his test object. The first campimeter developed by Aubert and Foerster (2) consisted of a flat surface with letters or figures arranged around a fixation point. The experiment was conducted in a dark room with the flash from an electrical discharge illuminating the surface. The next instrument developed by them consisted of a flat strip fastened to an upright in such a way that it could be rotated. From this later instrument Foerster (2) developed, in 1869, the curved arc instrument that is basically the perimeter so widely used today.

The measurement of peripheral vision on the arc perimeter is influ-

* This article is based on the author's thesis of the same title submitted to the faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, February, 1946. The thesis was directed by Dr. Joseph Tiffin.

enced by the size of the test object, its brightness, its color, its distance from the eyes, its background, the exposure time, and the amount of illumination. The subject introduces variability in his attention and the light adaptation of his eyes. Most of the attempts toward improvement of the perimeter have been to control and standardize these conditions.

Pascal (5) recently described an arc perimeter on which he uses a changeable fixation target and an illuminated test object. He uses a manually operated switch to flash the lights on and off. Mayer (4) described a test object, mounted on the arc of an ordinary perimeter, that is illuminated by a neon tube with a flash speed of .025 second. Burnham (1) developed a perimeter that he called a perihemisphere because it uses a hemispherical aluminum shell in which the test object can be located anywhere in the visual field. Color filters can be placed in the test object which is illuminated by a flashlight. The interior of the shell is painted white and brightly illuminated to present a uniform field.

In spite of the many attempts to improve the design for ease of operation and increased objectivity, two leading clinical perimetrists, Traquair (8) and Thomasson (7), prefer to use a simple adaptation of the original Foerster perimeter. Traquair (8) points out that perimetry is a highly subjective form of examination, an examination of the subject's sensations as described by himself in answer to questions put to him by the observer, and any success in the results is dependent more on the skill and knowledge of the experimenter than on the instrument used.

Previous Related Studies

Low (3) has reported a study dealing with peripheral vision and certain relationships between peripheral vision and other visual functions. He points out that a number of war-time accidents indicating faulty peripheral vision as a causative factor, prompted him to develop a reasonably short, accurate test of this visual function and to investigate the possibility of improving it by training. He used an ordinary arc perimeter with test objects of varying diameters.

The Apparatus and Test Procedure

Figure 1 shows the perimeter in use. It consists of a base, two swinging arms mounting lamps that contain the test objects, a stationary arm in the center mounting the fixation target and its lamp, protractors for measuring the angle of each swinging arm, enclosures, head rest, and the necessary wiring. The light control, an independent unit, consists of a double throw switch that turns the stimulus lamps on steady or into the flash timing circuit, a selector switch that turns the side test object

lamps on in various combinations with the center lamp and an electronically operated flash timing switch.

The experimental model operates only in the horizontal or temporal plane. It was made this way to simplify construction and yet allow for an evaluation of the basic concepts incorporated in the instrument.

A schematic drawing of the top view of the perimeter is shown in Figure 2. The examiner's control levers and protractors located under the center fixation target are not included in the drawing. All three



FIG. 1. The perimeter in use.

lamps are lighted with seven watt 110 volt candelabra bulbs. In the side or test object lamps the light is filtered and diffused through a dark filter and opal glass. On each side of the opal glass is a black opaque paper diaphragm with a centrally located $\frac{3}{16}$ inch diameter hole. These apertures on the opal glass serve as the test object when the light is turned on behind them. The opal glass is $17\frac{1}{2}$ inches from the eye. The test object size, given as a visual angle, is 37 minutes. This light source of low intensity is directed toward the eyes through one inch diameter tubes that are lined with lampblack to reduce reflection.

The center lamp has a one-half inch diameter aperture. A disc

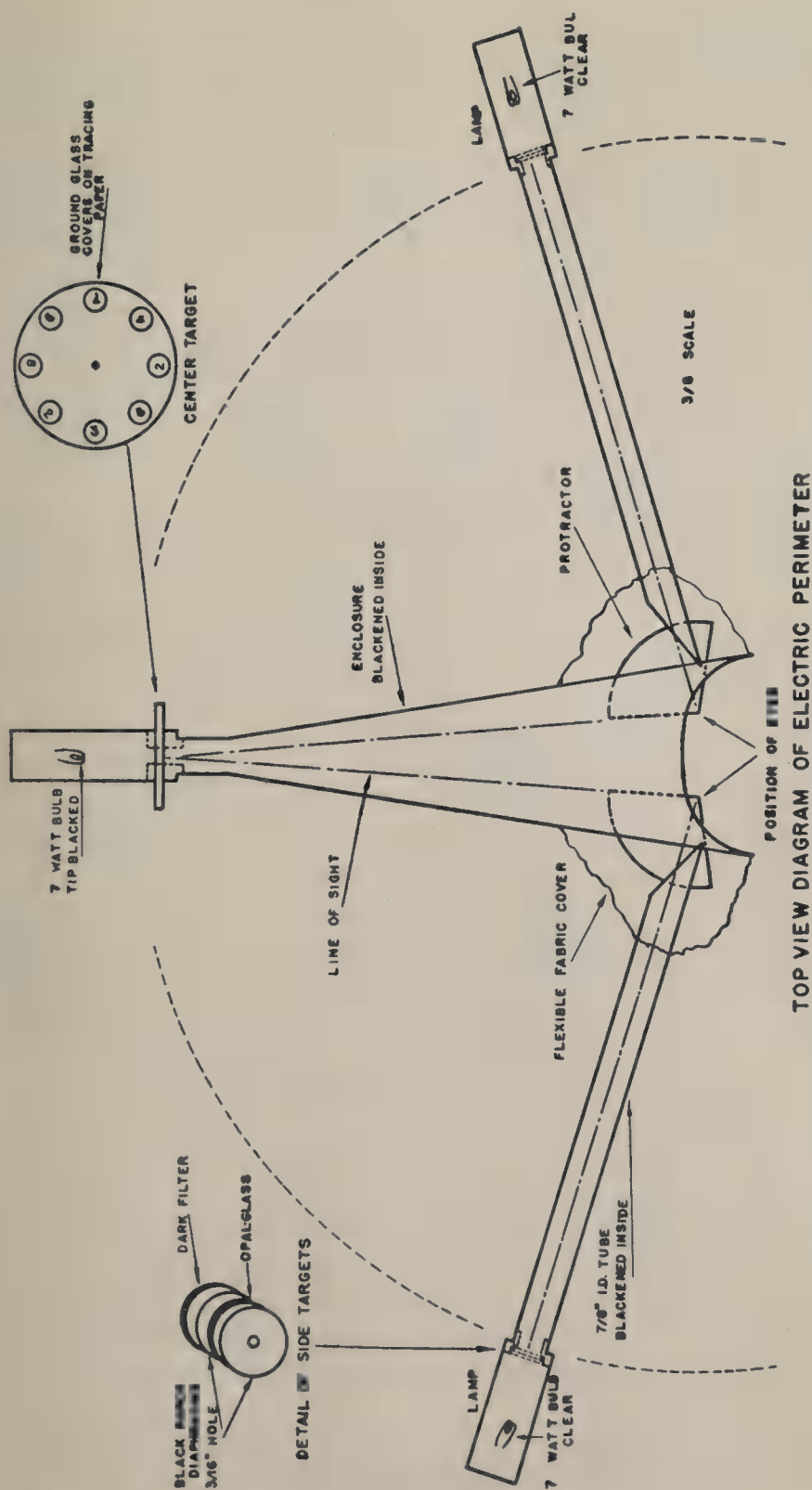


Fig. 2. Schematic drawing of the perimeter showing the relationships of the various parts.

located transversely in the tube in front of the center lamp mounts eight targets. These fixation targets are opaque one-quarter inch numerals on tracing paper held between ground glass covers. The disc is notched on its periphery so that the targets index accurately when the disc is rotated by hand. A small beam from a light under the center target illuminates the front of the center target aperture so that the subject can determine where to direct his attention between flashes of the light stimulus.

The headrest is part of the center metal enclosure. The enclosure is shaped so that when the subject's face is pressed slightly into the headrest, outside light is excluded, and the subject's eyes are positioned centrally in relation to each side lamp. The swinging arms are moved by levers on the protractors. The levers and protractors are located at the examiner's position in front of the instrument.

The circuit is wired so that the center lamp always lights. The test object lamps can be turned on with the center lamp in combinations of right-center, left-center, both-center, or neither-center. The electronic flash timer consists of a transformer, resistors, capacitors, an electronic tube and a relay. The wiring diagrams of these circuits can be found in a thesis ¹ located in the Purdue University Library.

The timer was set for a flash duration of one-tenth of a second. In preliminary trials it was found that this time gave sufficient exposure without allowing the subject time to shift his eyes.

The principal features built into the perimeter are: 1. Both eyes are tested simultaneously but the right eye cannot see the left field nor the left eye see the right field; 2. The subject must focus his attention on the center fixation target to read the number flashed. 3. The experimenter can determine whether the subject actually sees the test objects or is guessing; 4. The test can be given in less than ten minutes in its present form; 5. The test is not uncomfortable or fatiguing to the subject; 6. The intensity and duration of light stimulus can be carefully controlled; 7. The procedure and purpose is readily understood by the subject; 8. The field of vision is enclosed to reduce the effects of surrounding illumination; and 9. The instrument is readily portable.

The protractors are set to measure the side angle from straight ahead of the subject's eyes. Actually, when the eyes converge on the center fixation target the field angle for each eye is increased four degrees and twenty minutes. When comparing the results of a perimeter that measures one eye at a time with the results of this instrument this fact must be considered.

¹ McClure, John Allan, The development and standardization of a new type test for peripheral vision, Ph.D. Thesis, Purdue University Library, February, 1946.

The instrument was originally made with the center fixation target similar to the side test objects. To insure center fixation of the eyes and prevent falsification of responses, the test objects and center target first were made to flash successively, with the subject required to repeat the order in which the three lamps flashed. A motor driven brush contacting three adjustable contacts gave the sequence. Three tandem selector switches were wired into the contact circuit to give all six combinations of sequence for the three lamps. This method of presenting the stimulus was discarded because, as the test objects approached the peripheral vision threshold, the subject could not remember the sequence of flashes although he could clearly distinguish that a sequence had occurred. The test given in this manner seemed to be more a test of a special type of memory for perceptual experiences than a test of the field of vision. This method of administration was therefore abandoned.

Enclosure of the field of vision was found to be necessary because various external light sources affected the results when the instrument was not enclosed.

Test Procedure. The subject is seated on a stool that is adjusted to the correct height with the subject's eyes level with the headrest of the instrument. The subject is asked to remove his glasses if he wears them. The subject is told that the instrument is a perimeter for measuring how far to each side he can distinguish a dim flashing light while looking straight ahead. He is asked to press his face into the headrest so that his eyes are comfortably centered and so that no light enters around his face. The swinging arms are positioned at 45° from straight ahead. All lamps are turned on, after which the subject is asked what number he reads in the center target. He is then asked if he sees a small dim light on each side. Each side lamp is moved slightly while the subject is asked which one is moving. The examiner does not proceed until he is certain that the subject recognizes positively these side test objects. When testing those individuals with extremely narrow fields the arms are moved in closer than 45° . The purpose of the lighted aperture is explained briefly and demonstrated with a flash of light. The lamps are turned into the flash circuit. With each lamp set on 45° the combinations are explained while flashing center-right, center-left, center-both, and center-neither lamps. The subject is asked if he followed the combinations correctly. If necessary, the examiner again demonstrates and explains until this part is thoroughly understood. The subject is told to respond by telling what number he reads in the center target and which of the side lamps, if any, flash.

The selector switch is set to flash both side lamps. The examiner says, "Ready," just before he flashes the lights. If the response is cor-

rect the lamps are moved to 65° and again flashed. This large initial increase in the angle works well with the average subject in speeding up the testing procedure. During the practice trials both side lamps are flashed except when there is indecision on the part of the subject. In such cases the increments are smaller and more variations in the lamp combinations are given. If the response is correct on the 65° setting, the side lamps are moved to 75° , and then to 85° . From there on the increment is by five degree intervals. Both arms are always set at the same angle from straight ahead. When a setting is reached where the subject starts to give incorrect responses for either eye, or reports that he fails to see the test objects, three or four extra trials are given to be sure that the subject's threshold, on one or both of his eyes, has been passed. The arms are then brought forward five degrees to a smaller angle and four or five check trials are given. When the subject responds correctly on these practice trials the test trials are begun.

Table 1 shows a typical record sheet. The series of ten trials shown under *stimulus* is given and the response of the subject recorded under

Table 1
A Typical Individual Record Sheet

		Record Sheet									
Name: Mary Smith		Class: Psych. 1-B						Date: 1/21/46			
Angle		Stimulus									
		1	2	3	4	5	6	7	8	9	10
Left	Right	Both	Both	Both	Right	Left	Both	Both	Left	Right	Both
First Test											
85	85	B	B	B	R	L	B	B	L	R	B
90	90	B	B	B	R	L	B	B	R	R	B
95	95	R	R	R	R	R	R	R	O	R	R
100	100	O	O	O	O	O	O	O	O	O	O
Score—90L 95R											
Retest											
85	85	B	B	B	R	B	B	B	L	R	B
90	90	B	B	B	R	B	B	B	L	R	B
95	95	R	R	R	R	O	R	R	O	R	R
100	100	O	O	O	O	O	O	O	O	O	O
Score—90L 95R											

each stimulus trial. Of the ten trials given, eight involve an exposure of the stimulus on the right, and eight an exposure of the stimulus on the left. Thus, of the stimuli indicated across the top of Table 1, stimuli 1, 2, 3, 4, 6, 7, 9, and 10 are used in scoring the right eye (stimuli 5 and 8 having only *left* side exposures), whereas stimuli 1, 2, 3, 5, 6, 7, 8, and 10

are used in scoring the left eye (stimuli 4 and 9 having only *right* side exposures). Although the responses to the fixation target numbers are not recorded, consistent errors in calling the numbers are noted and the subject is encouraged to watch the target more carefully. If the subject can give correctly seven out of the eight responses for each eye the angle is increased by five degrees and the same series of trials is given. If he cannot give seven out of eight responses correctly for each eye the angle is diminished until a point is reached where seven out of the eight responses are given correctly for each eye. The angle is then increased in steps of five degrees, and the same series of ten trials is repeated until a point is reached where the subject states he cannot see the lights on either side or is consistently making errors so that the examiner is convinced the subject is guessing. All of the responses are recorded in the test trials.

In the typical series shown in Table 1, the score on the first test is 90° for the left eye and 95° for the right eye. The reason for these scores is that when an angle of 90° was used the only mistake made was on trial 8, which involved only the left eye, indicating that all eight trials involving the right eye were correctly reported. When the angle was increased to 95° , of the trials involving the left eye, trials 3, 5, 6, 7, 8, and 10 were reported as if there were no light on the left side, although actually a sight appeared in the left side in all of these trials. Since fewer than seven of the eight trials involving the left eye were correctly reported at 95° , the score for the left eye is recorded at 90° .

For the right eye, however, it will be noted that of the eight trials involving the right eye at 95° , every one resulted in a response indicating that the light on the right was seen whenever it was presented. Since at the next step, 100° , all responses were wrong, the score for the right eye was recorded at 95° .

Experimental Procedure

Subjects. Two hundred and two subjects enrolled in psychology courses were used as subjects. There were 96 males and 106 females. Their ages ranged from 16 to 40 with an average of 21.1 and a standard deviation of 8.37. Two subjects were scheduled for each half hour period.

Other Visual Skill Tests. The Bausch and Lomb Ortho-Rater was used to measure the visual skills of far vertical phoria, far lateral phoria, far acuity of both eyes, far acuity of right eye, far acuity of left eye, color vision, depth perception, near acuity of both eyes, near acuity of right eye, near acuity of left eye, near vertical phoria, and near lateral phoria.

Each subject was given the perimeter test first and then the Ortho-

Rater test followed by a retest on the perimeter. There was an approximate lapse of fifteen minutes between the first test and the retest with the perimeter. About one-third of the subjects were tested on days when the weather was cold and clear with bright sunshine on freshly fallen snow. Because of the close scheduling of subjects, only about five minutes could be allowed for light adaptation before the first test. The room lights were on during the testing.

Results

Method of Scoring. The record sheet of each subject was first scored to determine how many trials resulted in correct response for each eye at every angular setting marked. The largest angular setting was noted for each eye where seven out of eight responses involving that eye were correct. The largest angular setting was also noted for each eye where all of the responses involving that eye were correct. The sum of the right and left eye readings determined the included angle for each set of responses.

Reliability. The only published report found on the reliability of perimeter tests is that of Low (3), who found the reliability of his instru-

Table 2

Average Angular Thresholds and Standard Deviations of Left, Right, Both, and Total Fields for 7 out of 8 Correct Responses and 8 out of 8 Correct Responses with Correlations between First Test and Retests of Each

Criterion—7 of 8 N—202		Mean	S.D.	S.E. Mean	Diff.	S.E. Diff.	C.R.	<i>r</i>	S.E. _r
Left Eye	First Test	92.2	7.82	.550					
	Retest	94.7	6.70	.471	2.5	.331	7.55	.80	.025
Right Eye	First Test	93.3	7.26	.511					
	Retest	96.0	7.07	.497	2.7	.357	7.56	.75	.031
Included Angle	First Test	183.1	19.83	1.395					
	Retest	188.0	12.45	.876	4.9	.827	5.93	.83	.022
Criterion—8 of 8 N—169		Mean	S.D.	S.E. Mean	Diff.	S.E. Diff.	C.R.	<i>r</i>	S.E. _r
Left Eye	First Test	90.9	7.91	.608					
	Retest	93.9	7.25	.558	3.0	.390	7.69	.70	.030
Right Eye	First Test	91.7	7.73	.595					
	Retest	94.7	6.55	.504	3.0	.434	6.91	.70	.039
Included Angle	First Test	180.2	14.49	1.115					
	Retest	186.1	12.85	.988	5.9	.692	8.53	.79	.029

ment, an arc perimeter, to be .91. This test required from 40 to 60 minutes to administer and was of a clinical nature.

In the present investigation, correlations between the first test and the retest were obtained for the left eye, for the right eye, and for the included angle. The means and standard deviations of the scores as well as the correlations between the first test and the retest scores are given in Table 2.

The correlations are slightly higher when seven out of eight correct responses were used as the criterion than when eight out of eight were used. Because of its wider range, the measure of included angle gives the largest coefficient of reliability. The right eye test-retest correlation



FIG. 3. Frequency distribution of the total angular field.

gives the lowest coefficient of .75, the left eye a coefficient of .803 and the included angle a coefficient of .83.

The average scores on the retests were slightly, but significantly, higher than on the first tests. Evidently increased familiarity with the instrument and light adaptation account for the small increases in average scores. It was evident in the testing procedure that the subject usually erred only once in a series of ten trials if the test objects were located within his field of vision. A few check trials showed that when an error was made in this instance a second and immediate repetition of the stimulus combination usually resulted in a correct response.

Individual Differences. Figure 3 is a frequency distribution of the included angle of the visual field. This curve was plotted from the retest results using seven out of eight correct responses as the criterion in deter-

mining the score. The range is from 145° to 210°, or 65°. The mean is 188° with a standard deviation of 12.45°.

Figure 4 shows frequency distributions of the right eye field and the left eye field, again using retest results with seven out of eight responses correct as the criterion in determining the score. The mean angle for the right field, 96.0°, is slightly larger than the mean angle for the left field, 94.7°. This difference has a critical ratio of 4.2, indicating that the field of vision on the right is significantly, although only slightly, wider on the average than the field of vision on the left. The correlation between the size of the right and left fields found from the retest results, and used in

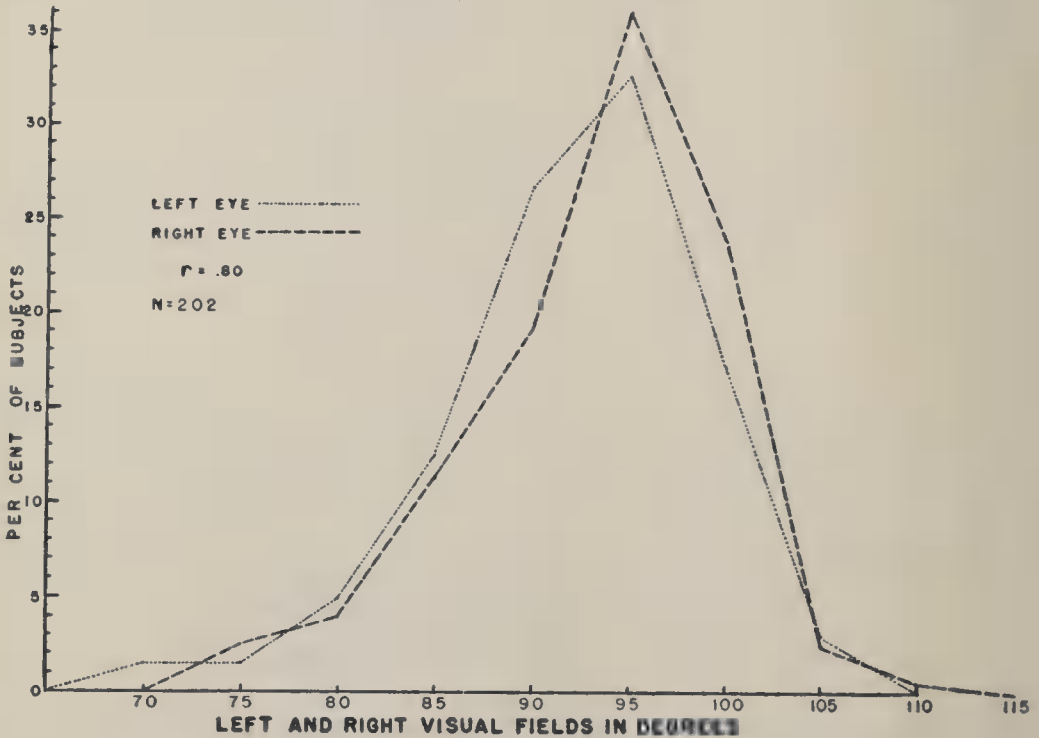


FIG. 4. Frequency distribution of the angular field of each eye.

determining the significance of the difference between the right and left fields, was .80, with an S.E. of .026.

Figure 5 shows the frequency distribution of the male and female included angle visual fields. The males had a mean included angle of 188.8°, S.D. = 8.85. The females had a mean included angle of 182.5°, S.D. = 14.1. The obtained difference of 6.3° between the mean included angle of the males and the females was 3.7 times as large as the S.E. of the difference, thus showing that the males have a slightly, but significantly, wider visual field than the females.

Relation to Other Visual Skill Tests. Table 3 lists the correlations

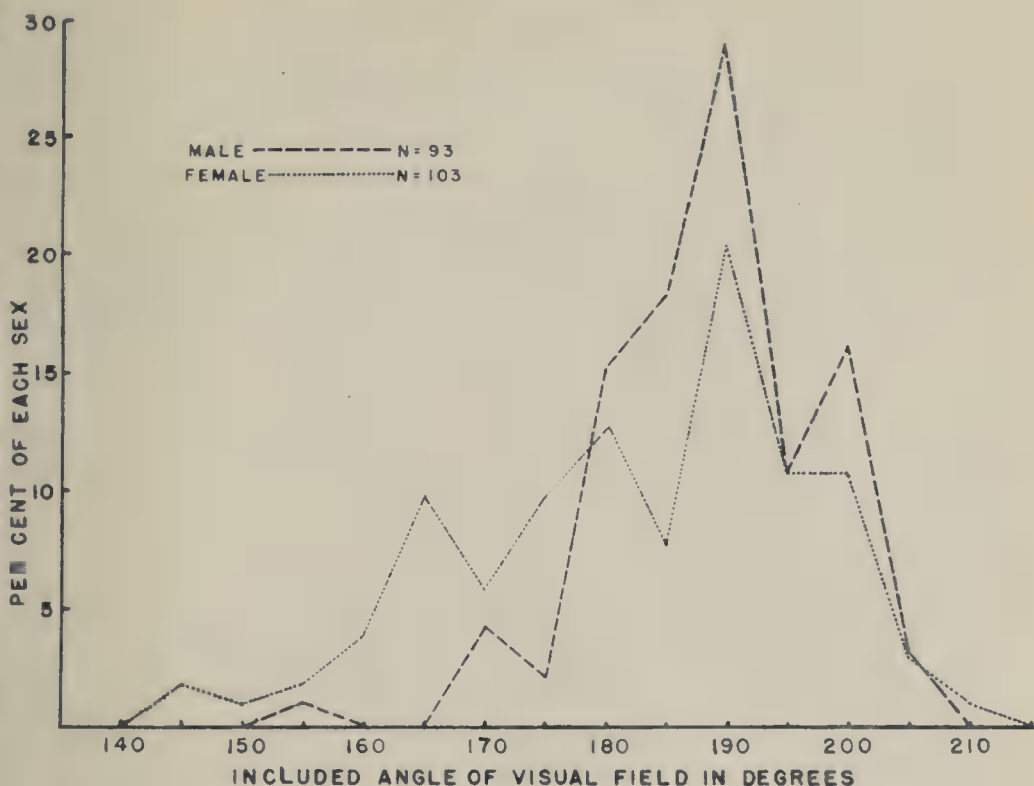


FIG. 5. Frequency distribution of the total angular field for males and females.

between the included angle as measured by the perimeter and the various visual skills measured by the Ortho-Rater. Table 4 shows the correlations between the near and far acuity of each eye with its visual field.

Table 3

Correlations of Ortho-Rater Tests with Perimeter Tests of Included Angle

	<i>r</i>	S.E. _r
Far Vertical Phoria	.00	.071
Far Lateral Phoria	-.01	.071
Far Acuity—Both Eyes	.12	.069
Far Acuity—Right Eye	.24	.066
Far Acuity—Left Eye	.16	.069
Far Acuity—Worse Eye	.26	.066
Depth Perception	.17	.068
Color Discrimination	.05	.070
Near Acuity—Both Eyes	.09	.069
Near Acuity—Right Eye	.12	.069
Near Acuity—Left Eye	.13	.069
Near Acuity—Worse Eye	.13	.069
Near Vertical Phoria	.01	.071
Near Lateral Phoria	.09	.069

Table 4
Correlation of Ortho-Rater Left and Right Eye Acuity Tests with
Perimeter Tests of Each Field

	<i>r</i>	S.E. _r
Far Acuity—Left Eye with Left Field	.18	.068
Near Acuity—Left Eye with Left Field	.13	.069
Far Acuity—Right Eye with Right Field	.24	.066
Near Acuity—Right Eye with Right Field	.11	.069

These obtained correlations and their standard errors indicate vision is not significantly related to these other visual skills, except, possibly, far acuity, and even in this case the relationship is very low.

The coefficient of correlation between age and peripheral vision was found to be .06, with an S.E. of .070.

Comparison with Other Investigations. Low (3) found no appreciable correlations with age, sex, central acuity, or color vision. He concludes that the correlation of .39 which he found between central acuity and peripheral field has no practical predictive value. This conclusion is in accord with the findings of the present study.

Summary and Conclusion

The perimeter described is adaptable for industrial and laboratory testing of peripheral vision limits because it tests rapidly and objectively, has satisfactory reliability, and can be operated successfully by an examiner without clinical training. The subject cannot falsify his responses because he must fixate both eyes on the center fixation target and the examiner, by controlling test object lamps flashes, can determine when the subject's responses are wrong.

The research has revealed a rather wide range of individual differences in the extent of the visual fields.

Peripheral vision, as measured by this instrument, has been found to be relatively independent of the visual skills of acuity, vertical and lateral phoria, depth perception and color discrimination.

Received April 27, 1946.

References

1. Burnham, R. S. A Perihemisphere for visual measurements. *J. exper. Psychol.*, 1940, 27, 333-336.
2. Foerster, R. Vorzeigung des Perimeter. *Elimische Monatsblätter für Augenheilkund*, Stuttgart, 1869, 7, 411-422.
3. Low, F. N. Studies on peripheral visual acuity. *Science*, 1943, 97, 586-587.

4. Mayer, L. H. Light stimuli of minimal durations as a means of perimetry. *Arch. Ophthal.*, Chicago, 1935, 14, 541-553.
5. Pascal, J. I. An improved perimeter-campimeter. *Arch. Ophthal.*, Chicago, 1937, 16, 103-105.
6. Purkinje, J. *Beobachtungen und Versuche zur Physiolo. der Sinne*, Prag, J. G. Calve, 1823.
7. Thomasson, A. H. A plea for greater uniformity in methods of field taking. *Arch. Ophthal.*, Chicago, 1934, 12, 21-32.
8. Traquair, H. M. *An introduction to clinical perimetry*. London: Henry Kimpton, Publisher, 1942.
9. Von Graefe, A. *Archiv für Ophthal.*, Berlin, 1855.
10. Young, T. Mechanism of the eye. *Royal Society of London, Philos. Trans.*, London, 1801.

Statistical Laboratory for Vision Tests at Purdue University

S. Edgar Wirt

Division of Applied Psychology, Purdue University

Employee tests of vision in a number of industrial plants are being tabulated and analyzed in a statistical laboratory in the Division of Applied Psychology at Purdue University. This Occupational Research Center uses modern electric punched card tabulating equipment to perform in a matter of minutes various types of statistical analyses that would require hours or days by other methods.

Vision tests are given to employees in these different plants by persons who have attended an intensive two-weeks training course, the Industrial Vision Institute, at Purdue. The test scores, marked on a self-scoring record form, along with other pertinent personnel data, are sent to Purdue. Here the data are transferred to punched cards on a machine that is operated like a typewriter. One punched card contains a complete transcript of the record of one employee, including his name, number, department, job, age, experience, vision test scores, etc. This transcript is in the form of holes punched into the card, and also in printed letters and numbers along the edge of the card. These cards for each job or department are tabulated, analyzed, and reported to the company as routine work of this Occupational Research Center.

Scattergrams

One of the most frequent types of statistical analysis performed in this laboratory involves the preparation of a series of scattergrams, plotting in turn each of 14 different measures of vision against a measure of success in job performance. Sometimes there are several different measures of job success (such as rate of production, earnings, quality of work, absences, merit ratings, accidents, etc.) each of which may be plotted against each of the vision tests. Each of these scattergrams is tabulated and automatically printed, line by line, completely in less than one minute for a hundred cases. Column headings are pre-printed on a special paper form. The tabulating machine does the rest—printing automatically on each row of the scattergram the vertical or *Y* category, the cell frequencies, total frequency, and sum of all *X* scores for that *Y* category. One scattergram is completed by running the cards once through the tabulator. A grand total, summing the values in each

column, requires a second run of the cards through the tabulator—again less than a minute for a hundred cases.

These scattergrams are the basis for the major work of the statistical laboratory, which is to evaluate the relations between visual requirements and job success for a particular job. The *degree* of relationship, as would be indicated by a coefficient of correlation, is not the most practical statement of this relationship. Instead it is necessary to determine for each vision test a critical score that may be recommended as a minimum or optimum for placement of an employee on a particular job. This can be done only on the basis of a scattergram. If a test is related to job performance, the better workmen will tend to fall predominately in one part of the range of the test while the poorer workmen tend to fall predominately in another part of the range.¹ The recommended critical score or "cut-off point" in the test range must help significantly and practically in differentiating between better and poorer workmen on the job.

Statistics for Large Groups

Another type of statistical analysis performed in the laboratory is the tabulation of vision test statistics based on large groups, including subjects on different jobs, in different plants, or in different communities. The purpose of such large scale tabulations is to establish norms on the tests for different groups and to compare frequency distributions and group statistics among different groups. This requires frequency distributions and scattergrams with large numbers of cases, classified by age, sex, job, length of experience on the job, section of the country, and so on.

The capacity of the tabulator for such large scale studies is tremendous. With a maximum of 17 categories in the variable plotted horizontally, the scattergram described above can show a frequency up to 10,000 in any one cell, and up to 100,000 in any row. With a maximum of 24 categories in the horizontal variable it can show a frequency up to 1,000 in any one cell and up to 10,000 in any row. This is in addition to the Y category designation and sum of scores in each row.

Correlations

A third type of statistical analysis performed in the bureau is the computation of correlation coefficients, particularly intercorrelations for multiple correlation of factor analysis. The coefficients are obtained by

¹ Tiffin, Joseph and Wirt, S. Edgar, Determining visual standards for industrial jobs by statistical methods. *Trans. Amer. Acad. Ophthal. and Otolar.*, 1945, Nov.-Dec., 72-93.

a tabulating process that yields directly the constants for computing the coefficients of correlations without going through the process of preparing scattergrams.² The constants obtained are those necessary to solve the formula:

$$r = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{N(\Sigma X^2) - (\Sigma X)^2} \sqrt{N(\Sigma Y^2) - (\Sigma Y)^2}}$$

which involves only raw scores for each intercorrelation. Intercorrelations between n variables required that the data cards be put through the tabulator n times, at a speed of 150 cards per minute. (For a 2-digit or 3-digit variable the cards may have to be put through two or three times instead of only once.) The number of variables and of cases that can be handled at one time is limited only by the size of the sums of the separate variables. The number of digits in sums of all the variables can not exceed 80. This permits 20 variables with sums of four digits each, 13 variables with sums of six digits each, and so on. For greater numbers the variables must be divided into two sections and intercorrelations plotted separately for sections aa, ab, and bb.

N (number of cases) and the Σ (sum) for each variable are read directly from the tabulated report. Each Σ^2 (sum of squared scores) and ΣXY (sum of cross products) is obtained by summing on an adding machine a column of figures produced in the tabulated report. The only remaining chore is to substitute in the formula and calculate the coefficients of correlation.

Item Analysis

A fourth type of statistical tabulation by machine is item analysis. In a battery of yes-no or multiple choice questions, each question must correlate with a criterion. This criterion may be the total score on the battery of questions, in which case the item must correlate with the total. Or the criterion may be some measure of success in another endeavor, in which case the item must contribute towards a prediction of success on this other measure. There are various short-cut methods³ for determining the value or effectiveness of an item in a battery of test questions, but each method is based on a frequency count of each possible answer on each item.

A modification of the tabulator method for producing scattergrams makes it possible to produce frequency distributions of all possible an-

² Warren, Richard, and Mendenhall, Robert M. *The Mendenhall-Warren-Hollerith correlation method*. New York, Columbia University, 1929.

³ Lawshe, C. H., A nomograph for estimating the validity of test items. *J. appl. Psychol.*, 1942, 26, 846-849.

swers simultaneously for 12 dual-choice questions, 8 three-choice questions, 6 four-choice questions, and so on—by putting a set of cards once through the tabulator. A different method provides an item count simultaneously on forty questions (25 questions for frequencies of 100 or more) by putting the cards three times through the machine for a battery of three-choice answers, five times for five-choice answers, and so on.

Listing

Another use of the tabulating machine is to list, or transcribe, all or part of the data punched on cards. (1) All cards in each set of data may be completely transcribed, a line for a card, as a permanent file record; any card that should become damaged or lost can be reproduced from this record. (2) Suitable descriptive headings are punched in cards and listed automatically at the beginning of each tabulated report. (3) The report to a cooperating company includes a list, produced on the tabulator, of names of employees whose visual skills do not meet the minimum requirements for the job—requirements that have been determined on a factual basis by statistical analysis. The company may notify these employees individually concerning their handicap and refer them to eye doctors for visual care, which in most instances can help them to meet requirements for the job.

The Laboratory

The work of the statistical laboratory is largely routine. Sets of data come in, are processed, and reported back to the companies that collected the data. Special projects are fitted into this routine schedule. Special projects may be special studies requested in connection with the vision tests, other studies on personnel tests of various types, studies on merit rating, job evaluation, research studies, and so on. The present volume of such statistical work requires a staff of two psychologists, two tabulating machine technicians, and a secretary.

The equipment includes calculators, files for data and punched cards, and the following International Business Machines tabulating equipment:

- 1 Alphabetic Accounting Machine with 80 counters, 88 print bars, 2 digit selectors, progressive totals, 20 comparing relays, 7 plugboards, and a speed of 80/150 cards per minute.
- 1 Counting Sorter with zone and digit selectors, and a speed of 400 cards per minute.
- 1 Alphabetic Printing Punch
- 1 Alphabetic Verifier
- 1 Reproducing Punch, with gang punch, 80 columns of comparison.

Pending delivery of the last two items, the Printing Punch has served as reproducing punch and gang punch, The Tabulator has served as verifier, the Counting Sorter has helped out on tabulation.

This statistical laboratory, with respect to equipment, personnel, and operating cost, is subsidized by the Bausch & Lomb Optical Company as part of a cooperative research project on occupational vision. It was evolved over a period of several years and has been doing routine research on vision in industry since July 1944. Another part of this project was the development by Bausch and Lomb of the Ortho-Rater, a battery of standardized precision vision tests validated for industry by Purdue University. These tests are produced by Bausch and Lomb. It is these standardized tests used in industry that are the basis for all the routine and some of the special statistical work of the laboratory. A third part of the project is the Industrial Vision Institute, an intensive two-weeks course at Purdue, given several times a year for representatives of industries that are using the Ortho-Rater and the Purdue Occupational Research Center in their own vision programs.

This method of research on occupational vision was developed in three stages. First, staff research men at Purdue developed this research approach to industrial problems of vision, made studies in various plants, and reported them back to the management. Second, representatives of industry were taught this procedure, with which they made studies in their own plants, using semi-mechanical methods of card tabulation. Third, the Occupational Research Center was developed to analyze data that has been gathered in these plants by management personnel. This has resulted in a very large and growing collection of research data on occupational vision at Purdue University.

Received May 22, 1946.

Motor Performance of Normal Young Men Maintained on Restricted Intakes of Vitamin B Complex *

Josef Brozek, Harold Guetzkow, Olaf Mickelsen, and Ancel Keys

The Laboratory of Physiological Hygiene, University of Minnesota

Clinical accounts of vitamin B-complex deficiencies emphasize general weakness, incoordination, and other indications of neuro-muscular deterioration. Such symptoms are not specific to vitamin deficiencies and are difficult to evaluate objectively in either clinical or survey studies.

In controlled experiments quantitative description of voluntary muscular performance is possible and has been utilized in several investigations in this Laboratory (14, 15, 16). Reliance can be placed on the results of psychomotor tests provided that (1) the methods and their applications are rigorously standardized, (2) stability of performance in the control (pre-experimental) period is secured by adequate training, (3) possible additional practice effects in the experimental group are accounted for by the use of a strictly comparable control group, and (4) all hints concerning the subjects' nutritional status or suggestions of symptoms are scrupulously avoided.

The B vitamins can be regarded as a group not only because of some common physical properties but because their natural distribution in diets frequently leads to rather parallel degrees of adequacy or deficiency for the several components of the complex. The work in this Laboratory has been devoted principally to thiamine, riboflavin, and niacin. Each of these enters into fundamental enzyme reactions of muscle and nerve metabolism (10). Restricted intakes of these B-complex vitamins might

* The work described in this paper was done under a contract, recommended by the Committee on Medical Research, between the Office of Scientific Research and Development and the Regents of the University of Minnesota. Important financial assistance was also provided by the Nutrition Foundation, Inc., the U. S. Cane Sugar Refiners' Association, N. Y., the Corn Industries Research Foundation, N. Y., Swift and Co., Chicago, the National Confectioners' Association, Chicago, the National Dairy Council, Chicago, and the Graduate Medical Research Fund, University of Minnesota. Merck and Co., Inc., provided a generous supply of pure vitamins. Most of the food materials were supplied by the Subsistence Branch, Office of the Quartermaster General, U. S. Army. We appreciate the constant assistance of the members of the Laboratory staff, particularly Dr. Austin Henschel, Dr. Henry Longstreet Taylor, and Miss Angie Mae Sturgeon. Dr. Howard Alexander, assisted by Messrs. Norris Schulz and Ralph Michener, handled the statistical analyses. Mr. Ersal Kindel constructed the test equipment.

act as limiting factors in human motor performance even at levels well above obvious clinical deficiencies.

This is a report on the psychomotor performance of normal young men maintained on diets restricted in the B vitamins but otherwise adequate. Thiamine, riboflavin, and niacin were regularly checked by analysis of the diet. The other members of the B-complex are assumed to have been supplied in parallel amounts since the diet was composed of varied natural foodstuffs. The general experimental program and procedures have been described in greater detail elsewhere (17).

Experimental Program

The subjects in this experiment were eight men, 20 to 32 years of age. They were emotionally and physically normal and free from signs or history of nutritional abnormalities which might have affected their vitamin requirements or ability to do moderately hard physical work. Before they served as volunteer subjects in this experiment the men had been for some months in Civilian Public Service Camps for conscientious objectors. In so far as could be ascertained the camp diets did not exceed the recommendations of the National Research Council (19). The men were under supervision throughout the experiment and no food other than that provided in the controlled diet was permitted at any time.

The experiment consisted of four consecutive parts:

- I) The "standardization" period of 41 days, including a period of 25 days of hard physical work reported previously (16).
- II) The prolonged "partial restriction" period of 161 days on a diet providing from one-third to two-thirds of the B-vitamin allowances recommended by the National Research Council (19).
- III) The "acute deficiency" period of 23 days on a diet almost devoid of B vitamins.
- IV) The "thiamine supplementation" period of 10 days in which the "acute deficiency" diet of the experimental subjects was abundantly supplemented with thiamine. The intakes of riboflavin and niacin were maintained on the same level as in the "acute deficiency."

During the standardization period the subjects were maintained on a uniform diet and followed an identical physical exercise schedule. In the psychomotor tests standardization involved training the subjects to a high and consistent level of performance.

Direct analyses of samples of meals eaten during the partial restriction period showed average intakes, per 1,000 cal., of 0.185 mg. of thiamine, 0.287 mg. of riboflavin, and 3.71 mg. of niacin. The four men who

Table 1

Design of the Experiment: B-vitamin Intake, in mg. per 24 Hours¹

d = diet, s = supplement, p = placebo

Nutritional Regimen	Dates	Source	Thia- mine	Ribo- flavin	Niacin	Subjects							
						G	Wi	Wa	T	S	Ja	N	Jo
Partial restriction	6/7-11/14	Diet	0.59	0.88	11.8	d	d	d	d	d	d	d	d
		Supplement	1.00	1.00	10.0	p	p	p	p	s	s	s	s
Acute deficiency	11/15-12/7 ²	Diet	0.03	0.05	0.40	d	d	d	d	d	d ³	d	d
		Supplement ⁴	1.5	1.5	10.8	p	p	s	s	p	p	s	s
Thiamine supplementation	12/8-12/12	Diet	0.03	0.05	0.40	d	d	d	d	d	d	d	d
		Supplement	10.00	0.00	0.00	s	s	p	p	s	p	p	p
		Supplement	1.5	1.5	10.8	p	p	s	s	p	p	s	s
	12/13-12/17	Diet	0.03	0.05	0.40	d	d	d	d	d	d	d	d
		Supplement	5.00	0.00	0.00	s	s	p	p	s	p	p	p
		Supplement	1.5	1.5	10.8	p	p	s	s	p	p	s	s
N. R. C. Recommended Dietary Allowances ⁵			1.5	2.0	15.0								

¹ The dietary regimen preceding partial restriction has been described in a separate report (14).² Saturation-test dose was given on 12/7, the 23rd day of deficiency, at 6:00 p.m. It contained 1 mg. thiamine, 1 mg. riboflavin, and 10 mg. niacin.³ Subject Ja developed an upper respiratory infection and was dropped from the experiment on Dec. 4th.⁴ In the period of "acute deficiency" and "thiamine supplementation" the control group received daily synthetic vitamins (1 mg. thiamine, 1 mg. riboflavin, 10 mg. niacin) plus 1.2 gr. of dried yeast containing approximately 0.5 mg. thiamine, 0.5 mg. riboflavin, and 0.8 mg. niacin.⁵ These daily amounts were recommended by the Food and Nutrition Board of the National Research Council (19) for moderately active men (3,000 Cal.) of average weight (154 lbs.).

served as experimental subjects received placebos, while the four control subjects were given supplements as indicated in Table 1. Each subject received additional daily supplements of 25 mg. of pyridoxine, 25 mg. of ascorbic acid, 5,000 I.U. of vitamin A, and 170 I.U. of vitamin D. The energy expenditure during this period was about 3,300 Cal. per day. The physical work on a motor-driven treadmill and the testing program were rigidly standardized. In the restricted group the average body weight, taken nude and before breakfast, was 141.8 lbs. during the standardization period and 140.5 lbs. at the end of partial restriction. The average weight of the supplemented subjects was 148.3 lbs. before the start of the experiment and 145.5 at the end of the partial restriction.

During the period of acute deficiency all men were fed a synthetic diet composed largely of cornstarch, sugar, vegetable shortening, and purified casein, to which minerals and vitamins were added to produce a "balanced" diet except for absence of the B complex. A pair of men from each of the two groups used in the partial restriction was supplemented during the acute deficiency. The distribution of subjects into restricted and supplemented groups is summarized in Table 2.

Table 2
Distribution of Subjects in the Different Phases of this Experiment

Subjects	Status in Partial Restriction	Status in Acute Deficiency
G, Wi	Restricted (R)	Deficient (RD)
S, Ja	Supplemented (S)	Deficient (SD)
Wa, T	Restricted (R)	Supplemented (RS)
N, Jo	Supplemented (S)	Supplemented (SS)

For the first two weeks of the acute deficiency, all subjects were able to do hard physical work which increased their caloric expenditure to about 4,000 Cal. per day. During the third week, two restricted-deficient subjects (RD) were unable to continue their treadmill work. One supplemented-deficient subject (SD) who became ill with an upper respiratory infection was dropped from the experiment. The three experimental subjects weighed, on the average, 128 lbs. at the start and 121 lbs. at the end of the acute deficiency. The average weights of the supplemented group were 152 lbs. and 153 lbs., respectively.

Methods

In most cases, the actual performance of a motor task is the best method we know for estimating the potential performance, the "work capacity," in that task. Performance in many types of work, particu-

larly when the component of coordination is prominently involved, can be predicted more accurately from a tryout performance than on the basis of measurable neurological, muscular, and other physiological or biochemical characteristics.



FIG. 1. A general view of the treadmill and of the equipment for the tests of speed and coordination.

The psychomotor battery used in this experiment included tests of strength, speed, and coordination. Standard dynamometers were used for measuring strength. In the test of speed of tapping a stylus was used to strike alternately two plates separated by a small barrier; the number of taps in the first and last 10 seconds of a half-minute tapping period was

recorded by an impulse counter. In the test of speed of gross body reaction the subject, while walking on the treadmill, was required to turn off the lighted one of three bulbs by bending over and striking the proper key. The keys were placed 18 inches from the floor of the treadmill. Reaction-time score is the average time of fifty reactions. The pattern tracing test involved eye-hand coordination; the speed of tracing the



FIG. 2. A subject performing the test of gross body reaction time.

pattern was kept constant and the scores consist of the number of contacts between the stylus and the side of the pattern, and of the total duration of these contacts. The ball-pipe test measured the speed of forearm and hand-movements involved in dropping ball-bearings through a one-foot conduit pipe (5).

All of these tests, with the exception of the two dynamometers, were performed while the subject was walking on a treadmill. In clinical B-complex deficiencies the peripheral neuropathy is reported to affect first

the lower extremities. Because of the possible development during acute deficiency of difficulties in walking some psychomotor measurements that did not involve walking were needed. Therefore, two additional tests which could be taken in a seated position were applied. In the test of toe reaction-time the subject reacted to auditory stimuli by flexing the big toe of the left foot and lightly pressing against a wooden board which stopped the timer. The Minnesota Rate of Manipulation test (28) was used to measure the speed of finger movements. The manipulation and toe reaction tests were applied only during the acute deficiency.



FIG. 3. A close-up of the pattern tracing board and the two tapping plates. The lights provide the "ready," "go," and "stop" signals for the tapping test.

Active cooperation on the part of the subjects is one of the necessary prerequisites for the valid use of psychomotor tests. This requirement taxes the skill of the experimenter in maintaining optimal motivation; this is particularly true when the testing sessions are spread out over a period of months. Data obtained in this and other experiments (15) demonstrate that a relatively constant, high level of motivation can be achieved.

In designing an experiment which would permit an evaluation of possible psychomotor deterioration, it is advantageous to bring performance in each test to a practice plateau before the start of the experimental period. Performance deterioration can then be measured without

practice effects entering as a complicating factor. During the standardization period our subjects were given 30 practice trials on the psychomotor tests used throughout this experiment.

The performance fluctuations from trial to trial in a number of the psychomotor tests are no larger than the variability of many standard physiological measurements. The percentage individual and group variabilities were computed for five plateau trials of the standardization period (Table 3).

Table 3
Trial-to-trial Fluctuation of Psychomotor Performance in the Standardization Period After Two Weeks of Intensive Practice *

Test	Individual Variability	Group Variability
1. Strength, hand-grip	4.0%	1.6%
2. Strength, back-lift	11.7%	2.2%
3. Speed, initial tapping	5.8%	0.9%
4. Speed, terminal tapping	4.3%	1.1%
5. Speed, gross body reaction time	5.3%	0.9%
6. Speed and coordination, ball-pipe	4.1%	1.0%
7. Coordination, pattern-tracing time of errors	17.0%	6.7%
8. Coordination, pattern-tracing number of errors	13.2%	5.5%

* The individual and group trial-to-trial variabilities were calculated according to the following formulae:

$$\text{Individual variability} = \sqrt{\frac{\sum_{i=1}^{t=8} \left[\frac{\sum_{t=1}^{t=5} (x - \bar{x})^2}{4} \right]}{8}}$$

$$\text{Group variability} = \sqrt{\frac{\sum_{t=1}^{t=5} (M - \bar{M})^2}{4}}$$

Both measures of variability are expressed as percentages of \bar{M} . The symbols are defined as follows:

- x = score of an individual
- \bar{x} = mean of his five scores obtained in trials on five successive days
- t = trials
- i = individuals
- M = mean of the eight individual scores for a trial
- \bar{M} = mean of the five trial-means

During the period of prolonged partial restriction the psychomotor functions were measured at intervals of two to three weeks. In the subsequent period of acute deficiency the psychomotor tests were given at weekly intervals. All measurements were made in duplicate.

At the end of partial restriction some equipment changes were necessary. For example, in the pattern-tracing test the tip of the stylus, which had been flattened out by frequent use, had to be replaced by a more rounded one; this made the test more difficult. The line starter for the reaction-time test was rebuilt so that the time followed more quickly the onset of the electrical light stimuli; this had the effect of lengthening the reaction-time. Because of such changes, the two experimental periods will be evaluated separately.

Results

The control scores obtained at the start of the prolonged partial restriction period are given in Table 4. The data for partial restriction are

Table 4

Control Scores at the Start of the Period of Partial Restriction

Note: Tests and units are as follows: 1) Hand-grip, in kg. 2) Back-lift, in kg. 3) and 4) Tapping, number of taps in the initial and terminal ten seconds of a half-minute work period. 5) Gross body reaction-time, in 1/120 sec. 6) Ball-pipe, number of passages of the ball through the conduit pipe in one minute. 7) Pattern tracing, duration of contact errors, in 1/120 sec. 8) Pattern tracing, number of contact errors.

Group	Restricted					Supplemented				
	Subjects					Subjects				
	G	Wi	Wa	T	Mean	N	S	Jo	Ja	Mean
1. Strength, hand-grip	47	43	63	66	55	63	58	53	57	58
2. Strength, back-lift	—	122	190	157	156	231	146	171	156	176
3. Speed, initial tapping	62	59	70	68	67	63	69	63	61	64
4. Speed, terminal tapping	61	64	64	60	62	57	64	57	53	58
5. Speed, gross body reaction time	40	48	45	48	45	42	41	44	45	43
6. Speed and coordination, ball-pipe	72	80	75	69	74	53	60	77	69	65
7. Coordination, pattern-tracing time of errors	226	152	95	173	162	183	176	198	239	199
8. Coordination, pattern-tracing number of errors	40	29	20	34	31	33	37	38	44	38

summarized in Table 5. The data obtained in acute deficiency, except for toe reaction-time and the manipulation test, are presented in Tables 6 and 7.

Table 5
Changes in Scores from Control to Terminal (153rd day) Performance in the Period of Partial Restriction

Group	Restricted					Supplemented					t-test
	Subjects					Subjects					
	G	Wi	Wa	T	Mean	N	S	Jo	Ja	Mean	
1. Strength, hand-grip	5	9	5	5	6.0	-1	-3	10	1	1.8	2.21
2. Strength, back-lift	—	9	-30	19	-0.7	7	-9	5	-2	0.2	0.07
3. Speed, initial tapping	3	3	5	2	3.2	12	2	12	5	7.8	2.72*
4. Speed, terminal tapping	-2	2	1	2	0.8	7	-3	7	0	2.8	1.17
5. Speed, gross body reaction time	3	6	5	10	6.0	11	5	2	18	9.0	1.24
6. Speed and coordination, ball-pipe	4	2	-2	5	2.2	7	10	12	1	7.5	2.90*
7. Coordination, pattern-tracing time of errors	-158	-63	-40	-70	-82.8	-102	-112	-94	-112	-105	1.34
8. Coordination, pattern-tracing number of errors	-23	-8	-9	-9	-12.2	-13	-21	-14	-18	-16.5	1.66

With 6 degrees of freedom the 5% level of *t* is 2.45; the 1% level is 3.71.

* Single asterisk indicates significance between 5% and 1% levels.

In general, there was only a slight difference in the performance of the experimental and the control group in the period of partial restriction. In two of the eight psychomotor tests continued practice produced larger improvement in the scores of the controls. This possible suggestion of slight impairment in the experimental group had its counterpart in the marginal disturbance of carbohydrate metabolism, reflected in a small

Table 6
Control Scores at the Beginning (4th day) of Acute Deficiency

Group	Deficient					Supplemented				
	RD	RD	SD	SD		RS	RS	SS	SS	
Subgroup	Subjects					Subjects				
Test	G	Wi	S	(Ja)*	Mean	Wa	T	N	Jo	Mean
1. Strength, hand-grip	51	56	57	(58)	55	65	75	63	62	66
2. Strength, back-lift	—	136	140	(158)	138	153	176	239	175	186
3. Speed, initial tapping	61	73	66	(62)	67	71	66	65	72	69
4. Speed, terminal tapping	57	64	56	(50)	59	59	58	61	60	60
5. Speed, gross body reaction time	57	69	57	(87)	61	68	81	77	65	73
6. Speed and coordination, ball-pipe	73	85	76	(67)	78	77	75	59	89	75
7. Coordination, pattern-tracing time of errors	126	175	148	(203)	150	86	171	215	160	158
8. Coordination, pattern-tracing number of errors	21	31	25	(34)	26	14	29	36	32	28

* Ja was dropped from the experiment before its completion because of a respiratory infection.

increase in the blood pyruvate level (17). It also paralleled small changes obtained in the Rorschach.

During acute deficiency the psychomotor performance deteriorated markedly in the experimental group. These functions were among those aspects of fitness which showed an early deterioration, immediately following the gastrointestinal disturbances. The psychomotor tests were more sensitive to the "stress" of vitamin deficiency than many of the metabolic, neurological, and cardiovascular tests (17).

Terminal supplementation with thiamine alone for 10 days led to recovery in the tests of speed and of coordination. The small decrease in grip strength which appeared in the last week of deficiency was still present at the 10th day of supplementation.

These results will be discussed in detail for each test separately.

Table 7

Changes in Scores from Initial (4th day) to Terminal (23rd day) Performance in Period of Acute Deficiency

Group	Deficient				Supplemented					
Subgroup	RD	RD	SD		RS	RS	SS	SS		
	Subjects				Subjects					
Test	G	Wi	S	Mean	Wa	T	N	Jo	Mean	t-test
1. Strength, hand-grip	-2	-3	-3	-2.7	4	-2	0	3	1.2	2.37
2. Strength, back-lift	—	-20	21	.5	-16	-7	-3	16	-2.5	0.18
3. Speed, , initial tapping	-7	-12	1	-6.0	3	2	7	-4	2.0	1.93
4. Speed, terminal tapping	-8	-6	1	-4.3	5	1	4	-2	2.0	2.15
5. Speed, gross body reaction time	18	66	26	36.7	2	2	-3	11	3.0	2.61*
6. Speed and coordination, ball-pipe	-7	-7	-9	-7.7	2	3	-3	8	2.5	3.77*
7. Coordination, pattern-tracing time of errors	118	106	109	111.0	-32	-66	-97	-49	-61.0	10.31**
8. Coordination, pattern-tracing number of errors	24	11	13	16.0	-2	-12	-17	-12	-10.8	5.32**

With 5 degrees of freedom, the 5% level of t is 2.57; the 1% level is 4.03.

* Single asterisk indicates significance between 5% and 1% levels.

** Double asterisks indicate significance at better than 1% level.

Strength. Simple strength, measured by hand-grip and back-lift dynamometers, was in general remarkably stable. During the prolonged period of partial restriction there was no deterioration in either group; in fact the restricted group showed a slight but statistically not significant increase in hand grip. In the acute deficiency period the experimental group exhibited a very slight decrease in grip strength.

Speed: Finger Movements. The tests of speed of finger movements

was practiced at the end of the period of partial restriction but was experimentally used only in acute deficiency. The deficient group showed a slight decrease in efficiency, with a score of 87.4 at the fifth day and 84.6 on the twenty-third day of acute deficiency; the score is the number of discs turned over in the last minute of a five-minute work period. The control group continued to improve, having an initial score of 89.6 and the terminal score of 94.2. The difference between the two groups in the change from the beginning to the end of acute deficiency was not statistically significant, $t = 2.14$ (t at 5% level = 2.57).

Speed: Tapping. There was no loss of motor speed, as determined by performance on the two-plate tapping test, during prolonged partial restriction. There were changes in the positive direction, both groups slightly improving as a result of continued practice. The mean gain of the supplemented group was statistically significantly higher for scores obtained in the initial ten seconds of the half-minute work period, but the gains in the terminal ten second scores did not differentiate the two groups. Throughout the period of acute deficiency, the performance of the deficient subject *S* who was supplemented during the preceding 5 months remained unchanged. The two restricted-deficient subjects exhibited a deterioration of performance in both tapping scores.

Speed: Gross Body Reaction. The average time involved in selecting and striking a proper telegraph key while walking on the treadmill increased slightly during prolonged restriction in both groups. This is attributable to increased slippage in the brake mechanism of the timer. There was no significant difference between the reaction-time scores of the restricted and the supplemented group. In the subsequent period of acute deficiency the performance of the deficient group deteriorated markedly. The difference between the increases of the two groups in reaction times for the first day and the twenty-third day was statistically significant at the 5% level. Biologically, this increase in reaction time was more important than the 5% level of significance might indicate. In terms of percentages, the deficient group exhibited a 60% decrease in this aspect of fitness between the fourth and the twenty-third day of acute deficiency, as compared with a change of only 4% in the supplemented group.

Speed: Toe Reaction. In the other reaction-time test, the subject reacted with his big toe to an auditory stimulus. The toe reaction-time score was the average of fifty reactions. The deficient subjects averaged 43 on the third day and 58 on the twenty-third day of the acute deficiency period; the comparable mean scores for the supplemented group were 44 and 45, all scores being given in 1/120 second. The difference between the mean change from start to end of acute deficiency was significant at

better than the 5% level, $t = 3.07$ (t at 5% level = 2.57). The toe-reaction test did not appear to be more sensitive to the experimentally produced B-complex deficiency than the test of the gross body reaction-time.

Speed and Coordination: Ball-pipe. In the ball-pipe test the scores in the prolonged partial restriction showed slight effects of continued practice; this was more pronounced in the supplemented group than in the restricted group. In the subsequent period of acute deficiency there was definite evidence of deterioration in the unsupplemented group. The difference between the average change in the experimental and the control group from the start to the end of the deficient diet approaches the 1% level of statistical significance.

Coordination: Pattern Tracing. In the pattern-tracing test the scores continued to improve during the whole length of the prolonged partial restriction period. This was largely an artifact due to the gradual flattening of the point of the tracing stylus which made the task easier. Both the restricted and the supplemented group reduced the initial scores in approximately the same ratio. This ratio was 0.49 and 0.47 for the time-score of the restricted and the supplemented groups respectively, and 0.61 and 0.58 for the number-score. As has been stated, a new stylus-point was used at the beginning of the acute deficiency. The supplemented group adapted to this change whereas the performance of the deficient group showed striking deterioration. The difference between the experimental and the control group was statistically significant beyond the 1% level.

The change from initial score to terminal score was used as a measure of the overall effect of the experimental regimens. The t -tests presented in Tables 5 and 7 were based on these differences. In doing this we ignored the scores obtained during the period between the initial and the terminal testing session. Both the analysis of variance and a regression analysis of trends was applied to the full data. These analyses gave substantially the same results as the t -tests applied to the initial-terminal changes.

Discussion

In general, the problem of the relationship between vitamin intake and level of motor fitness comprises two rather distinct questions: performance on *supplemented* diets and performance on *restricted* diets. All acceptable studies of the first category are in agreement that extra vitamin supplementation of ordinary "good" diets, i.e., diets not considered really deficient, does not lead to improved muscular performance in relatively normal persons (8, 12, 13, 22).

In disturbed or special metabolic conditions vitamin supplementation may be beneficial. Such a positive effect has been reported for senile patients whose diet was heavily supplemented with B vitamins (23). In the study cited it appeared that there was at least a temporary response in psychomotor speed, coordination, and strength, but there is a possibility that the vitamin supplementation only helped to improve an otherwise poor hospital diet.

The question is more complex with restricted diets. If the restriction is sufficiently severe to produce obvious clinical deficiency, it is agreed that motor performance will be impaired, though the precise degree and nature of the impairment has not been characterized quantitatively. The effect of any restriction may be dependent upon many factors, such as previous diet, duration of the experimental dietary regimen, level of activity, and climate, as well as peculiarity of the subjects and the particular functions measured. The present confused state of the vitamin "requirements" problem reflects insufficient attention to these secondary factors. The present paper is concerned only with the capacity for relatively brief psychomotor performance of normal young men leading a moderately active life in a temperate climate. There are no published data directly comparable to those presented here.

Experiments carried out on animals provide little information bearing directly on those aspects of performance which were studied in the present experiment. The animal experiments deal with such aspects of "fitness" as prolonged work of intact animals (18), spontaneous activity (4, 9), and disturbance of locomotion and posture (7, 20).

Few of the studies on man have utilized adequate techniques to determine the effects of vitamin restriction on psychomotor performance. Reports of changes in total work capacity developing during subsistence on restricted intakes of the B vitamins emphasize decreased endurance in hard work (2, 3, 11). Such changes probably are dependent to a large extent on cardiovascular functions; in any case the reports cited provide little or no data on the more purely neuromuscular functions under present consideration. Simple clinical observations of the deterioration in motor behavior in non-standardized situations are valuable as clues for experimental work, but are difficult to evaluate (25, 26). If performance deteriorates, and that is by no means always clear, it is uncertain to what extent this is a result of changes in neuromuscular functions or of uncontrolled changes in motivation.

Williams and his colleagues expressed the opinion that measurements of performance capacity "are exceedingly difficult to make, for they involve not only the ability but also willingness to perform, and willingness is lost early in thiamine deficiency" (27, p. 72). It is true that a

performance score cannot be accurately divided into "capacity" and "willingness" components. However, when performance is characterized in quantitative terms and sufficient intra-individual and inter-individual controls are provided, this difficulty does not prevent experimental work in this area. It should be clear that, if "willingness" were more easily lost than capacity, we should be safe in concluding the absence of change in capacity when there is no change in total performance. This was the case in the period of partial restriction. One would expect that an unpleasant task, repeated month after month, would more likely be susceptible to changes ascribable to decreased "willingness" than a task equally often repeated but well liked. Note that performance on the back-lift dynamometer, a "back-breaking" task which was disliked by a majority of the subjects, did not show decrease over the prolonged partial restriction. There was no essential difference in this test as compared with the hand-grip dynamometer, although the men enjoyed the latter task. In the acute deficiency the performance in the majority of *intellective* tests, requiring a good deal of concentration and effort, remained unchanged; this demonstrates that "willingness" was not fundamentally affected (17). Yet, there was deterioration of *psychomotor* performance.

The present studies represent a development and continuation of previous investigations in this Laboratory. The studies of 1941-42 disclosed no deterioration in any of the functions when moderately active men were maintained for 10 weeks on a diet providing only 0.23 mg. of thiamine per 1,000 Cal. (14). A battery of psychomotor tests was applied in experiments involving subsistence for 152 days on a diet providing only 0.31 mg. of riboflavin per 1,000 Cal. Again, there was no functional deterioration observed (15). Finally, there was no deterioration in normal young men maintained for 14 days at hard work (4,600-4,800 Cal. daily) with a B vitamin intake per 1,000 Cal. at about one-fourth the National Research Council Recommended Daily Allowances (16).

Such results might suggest that the particular tests used in the present experiment are insensitive to "stress." This is disproved by positive results in other conditions such as fasting (24) and bed rest (6). The changes in psychomotor performance obtained in the acute deficiency period of the present experiment afford further evidence that these methods are not insensitive when dietary vitamin insufficiency is present.

The acute phase of the present experiment was designed to produce a physiological B-vitamin deficiency in a relatively short time by maintaining the experimental subjects on a diet extremely low in thiamine, riboflavin and niacin, and by a simultaneous high caloric output. Anima,

experiments have indicated that physical exercise speeds up the onset of deficiency symptoms in animals receiving a diet severely restricted in thiamine. In Guerrant and Dutcher's experiments this period was 36 days for the "forced exercise" group, and 44 for the "confined" group (9). In the acute deficiency period of the present experiment the treadmill work was intensified and the total daily caloric expenditure increased to about 4,000 Cal.

Studies of psychomotor performance by the techniques of factorial analysis indicate that motor skills are very specific; scores obtained in various motor tasks exhibit, on the average, positive but very low correlations (21). It is interesting that in acute deficiency the deterioration tended to affect all psychomotor functions. The degree of deterioration, however, varies. Since the number of subjects in the psychomotor tests was identical, the "degrees of freedom" are the same and the *t*-values for the different psychomotor tests are directly comparable. The relative sensitivity of the different tests to this stress is indicated indirectly by the *t*-value of the difference between the deficient and supplemented group. In this respect, deterioration in the pattern tracing test was most outstanding.

A more direct approach to determine the relative deterioration of the various functions is to express the mean changes from initial to terminal performance level of the deficient group in terms of the percentage of the average scores at the start of acute deficiency. This method of evaluating the deteriorative changes points in the same direction as the comparison of the levels of statistical significance (Table 8). However, use of percentages in comparing changes which occurred in various functions involves assumptions whose effect it is difficult to evaluate (1).

From the biological point of view, the most meaningful approach to the evaluation of the magnitude of changes which occurred in acute deficiency would be to relate them to the total range of performance in a given task in the general population. This is not possible at the present time. It would require use of rigorously standardized test techniques which in turn would make possible pooling of the data obtained by the various laboratories. In the area of psychomotor performance this lack of standardization is more glaring than in most areas of psychological and physiological testing.

It might be asked whether the decrease in performance in the period of acute deficiency of thiamine, riboflavin, and niacin was not due simply to partial starvation. In answering this question, two types of evidence should be considered. First, when Kniazuk and Molitor (18) reduced the food intake of their vitamin supplemented rats to that of the deficient group, they found no significant reduction in work performance. The

differences in the rate of recovery in our experiment also provide a good argument for a relative independence of this small and gradual loss of body weight and performance deterioration: the performance of the subjects in the deficient group recovered strikingly in the supplementation period whereas their weight increased only slightly.

The psychomotor tests used in our experiment were all of short duration. It is possible that more prolonged tests would have shown even more pronouncedly the effects of the vitamin deficiency. However, it is very difficult to devise satisfactory laboratory tests of work endurance

Table 8

Sensitivity of Psychomotor Tests to the Stress of Acute Vitamin B Deficiency

Test	Significance of <i>t</i> -test	Percentage Decrement in Deficient Group
Speed of finger movements	not significant	3.4
Strength, hand-grip	not significant	4.9
Speed, terminal tapping	not significant	7.3
Speed, initial tapping	not significant	9.0
Speed and coordination, ball-pipe	5% level	9.9
Speed, toe reaction time	5% level	28.7
Speed, gross body reaction time	5% level	60.2
Coordination, pattern-tracing number of errors	1% level	61.5
Coordination, pattern-tracing time of errors	1% level	72.0

of humans unless we use tasks which appreciably increase the energy consumption or lead to exhaustion of local muscle groups in a short time. The lack of adequate endurance tests is particularly unfortunate when we are interested in the application of the research findings to industrial nutrition. In modern industry there are few jobs producing physical exhaustion fatigue. Prolonged, repetitive work of moderate intensity is characteristic of the overwhelming majority of industrial jobs. In recent years the most successful approach to the problem has been made in experimental aviation medicine. The use of "miniature job situations" in the study of endurance in pilots should be paralleled in research on industrial physiology by the use of miniature industrial plants. This would extend the validity and practical importance of strictly laboratory studies.

Summary and Conclusions

1. The relationship between intake of B vitamins, particularly thiamine, and psychomotor performance was studied. This was one aspect of a comprehensive investigation of the biochemical, physiological, and psychological aspects of "fitness" as related to the vitamins of the B complex.

2. Eight "normal" men, 20 to 32 years of age, served as subjects. They were maintained for 161 days on a diet providing, on the average, 0.185 mg. of thiamine, 0.287 mg. of riboflavin and 3.71 mg. of niacin, per 1,000 Cal. The physical exercise was such that a daily intake of approximately 3,300 Cal. just maintained body weight. Four men received a daily supplement of 1.0 mg. thiamine, 1.0 mg. riboflavin and 10 mg. niacin, while the other four received placebos.

3. The period of partial restriction was followed by 23 days on a diet practically free of these vitamins. The subjects were re-grouped into the following four pairs: Restricted-deficient, restricted-supplemented, supplemented-deficient, supplemented-supplemented. The experiment ended with a 10 day period of thiamine supplementation.

4. In the period of partial restriction there was no actual deterioration in any of the psychomotor measurements, including two strength tests (hand-grip and back-lift), speed of small hand movements (tapping), gross body reaction time, manual speed-and-coordination (ball-pipe test), and precise coordination (pattern tracing). However, the scores in the initial 10 sec. of tapping and in the ball-pipe test showed larger practice increments in the supplemented than in the restricted group; the difference was small but statistically significant at the 5% level.

5. In the period of acute deficiency there was evidence of marked deterioration. The difference between the experimental and the control group was significant at the 5% level in the tests of gross body reaction-time, toe reaction-time, and ball-pipe, and at the 1% level in pattern-tracing. The decrease in two-plate tapping and in the test of the speed of finger movements was not statistically significant. There was a very slight decrease in grip strength, approaching the 5% level of significance. The measured deterioration of psychomotor performance, present as soon as the 11th day, was one of the early symptoms of deficiency.

6. Performance tended to return to a normal level after 10 days of liberal supplementation of the experimental diet with synthetic thiamine.

Received July 12, 1945.

References

1. Anastasi, A. Practice and variability. *Psychol. Monogr.*, 1934, 45, No. 204.
2. Archdeacon, J. W., and Murlin, J. R. The effect of thiamine depletion and restoration on muscular efficiency and endurance. *J. Nutrition*, 1944, 28, 241-254.

3. Barborka, C. J., Foltz, E. E., and Ivy, A. C. Relationship between vitamin B complex intake and work output in trained subjects. *J.A.M.A.*, 1943, **122**, 717-720.
4. Bloomfield, A. L., and Tainter, M. L. The effects of vitamin B deprivation on spontaneous activity of the rat. *J. lab. clin. Med.*, 1943, **28**, 1680-1690.
5. Brozek, J. A new group test of manual skill. *J. gen. Psychol.*, 1944, **31**, 125-128.
6. Brozek, J., Guetzkow, H., and Keys, A. Changes in psychomotor performance in bed-rest. *Fed. Proc. Amer. Soc. exp. Biol.*, 1945, **4**, 10.
7. Everett, G. M. Observations on the behavior and neurophysiology of acute thiamine deficient cats. *Am. J. Physiol.*, 1944, **141**, 439-448.
8. Foltz, E. E., Ivy, A. C., and Barborka, C. J. Influence of components of the vitamin B complex on recovery from fatigue. *J. lab. clin. Med.*, 1942, **27**, 1396-1399.
9. Guerrant, N. B., and Dutcher, A. A. The influence of exercise on the growing rat in presence and absence of vitamin B₁. *J. Nutrition*, 1940, **20**, 589-598.
10. Himwich, H. E. The role of the vitamins in brain metabolism. *Res. Publ. Ass. nerv. ment. Dis.*, 1943, **22**, 33-41.
11. Johnson, R. E., Darling, R. C., Forbes, W. H., Brouha, L., Egana, E., and Graybiel, A. Effects of a diet deficient in part of the vitamin B complex upon men doing manual labor. *J. Nutrition*, 1942, **24**, 585-596.
12. Karpovich, P. V., and Millman, N. Vitamin B₁ and endurance, *New Eng. J. Med.*, 1942, **226**, 881-882.
13. Keys, A., and Henschel, A. Vitamin supplementation of U. S. Army rations in relation to fatigue and the ability to do muscular work. *J. Nutrition*, 1942, **23**, 259-269.
14. Keys, A., Henschel, A., Mickelsen, O., and Brozek, J. The performance of normal young men on controlled thiamine intakes. *J. Nutrition*, 1943, **28**, 399-415.
15. Keys, A., Henschel, A., Mickelsen, O., Brozek, J., and Crawford, J. H. Physiological and biochemical functions in normal young men on a diet restricted in riboflavin. *J. Nutrition*, 1944, **27**, 165-178.
16. Keys, A., Henschel, A., Taylor, H. L., Mickelsen, O., and Brozek, J. Absence of rapid deterioration in men doing hard physical work on a restricted intake of vitamins of the B complex. *J. Nutrition*, 1944, **27**, 485-496.
17. Keys, A., Henschel, A., Taylor, H. L., Mickelsen, O., and Brozek, J. Experimental studies on man with a restricted intake of the B vitamins. *Am. J. Physiol.*, 1945, **144**, 5-42.
18. Kniazuk, M., and Molitor, H. The influence of thiamine deficiency on work performance in rats. *J. Pharm. exp. Therap.*, 1944, **80**, 362-372.
19. National Research Council. Recommended dietary allowances, revised 1945. Reprint and circular series No. 122, 1945.
20. Prickett, C. O. The effect of a deficiency of vitamin B₁ upon the central and peripheral nervous systems of the rat. *Am. J. Physiol.*, 1934, **107**, 459-470.
21. Seashore, R. H., Buxton, C. E., and McCollom, I. N. Multiple factorial analysis of fine motor skills. *Am. J. Psychol.*, 1940, **53**, 251-259.
22. Simonson, E., Enzer, N., Baer, A., and Braun, R. Influence of vitamin B (complex) surplus on capacity for muscular and mental work. *J. Indust. Hyg. Toxicol.*, 1942, **24**, 83-90.
23. Stephenson, W., Penton, C., and Korenchevsky, V. Some effects of vitamin B and C on senile patients. *British med. J.*, 1941, **2**, 839-844.
24. Taylor, H. L., Brozek, J., Henschel, A., Mickelsen, O., and Keys, A. The effect of successive fasts on the ability of men to withstand fasting during hard work. *Am. J. Physiol.*, 1945, **143**, 148-154.

25. Williams, R. D., and Mason, H. L. Further observations on induced thiamine (vitamin B₁) deficiency and thiamine requirements of man: Preliminary reports. *Proc. Staff Meet. Mayo Clinic*, 1941, 16, 433-438.
26. Williams, R. D., Mason, H. L., Power, M. H., and Wilder, R. M. Induced thiamine (vitamin B₁) deficiency in man: Relation of depletion of thiamine to development of biochemical defect and of polyneuropathy. *Arch. int. Med.*, 1943, 71, 38-53.
27. Williams, R. D., Mason, H. L., and Wilder, R. M. Minimum daily requirements of thiamine of man. *J. Nutrition*, 25, 1943, 71-97.
28. Ziegler, W. A. *Minnesota rate of manipulation test*. Minneapolis, Minn.: Educ. Test Bureau, 1939, 9 pp.

Standardization of a Test of Hand Strength *

M. Bruce Fisher

Fresno State College, California

and

James E. Birren

Northwestern University

In the course of developing a battery of sensorimotor tests for the assessment of the efficiency of naval personnel under various conditions of stress, it was believed desirable to include the measurement of some function dependent on muscular strength. The extensive use of the hand dynamometer over the past hundred years (7, 10) suggested this device as the test apparatus. Furthermore, it is easily adaptable to a wide variety of testing circumstances, both in the field and in the laboratory. High reliability was desired since the proposed experiments in which the test would be used were to involve small numbers of subjects. An additional desired characteristic was that the test should involve a small amount of fatiguing work—at least more than usually occurs when the best of two or three grips is taken as the score. Dunlap's endurance dynamometer test (2) meets this second requirement but he reports a low retest reliability. This finding of low reliability was confirmed in a study at the Naval Medical Research Institute. Other exploratory work led to the selection, for further standardization, of the procedure to be described.

Procedure

Test Administration. Each Smedley dynamometer was modified by affixing, over the regular scale, one which was divided into 3-kg. units, numbered in order. Thirds of a scale division (1 kg.) were indicated by smaller divisions. Although the major purpose of this scale is to facilitate easy reading, its placement permits the correction of any error in the original scale. In four dynamometers calibrated for this study, the

* This report was prepared when the writers were on active duty in the U. S. Naval Reserve at the Naval Medical Research Institute, National Naval Medical Center, Bethesda, Md. The opinions expressed are those of the writers and are not to be construed as reflecting the policies of the Navy Department.

calibration curves were approximately parallel straight lines, but the true zero points varied over a 3-kg. range on the original scales.

In a test, each *S* adjusted the length of the movable stirrup to fit his hand, being warned not to make this adjustment too short. Each *S* was also instructed to grasp the dynamometer with the fixed stirrup bearing across the heel (hypothenar eminence) of his hand so that all four fingers would have a chance to work, and so that painful pressure on the soft tissues between thumb and index finger would be avoided. To assist in the maintenance of consistent performance and facilitate reading of the scale between pulls, *S* always gripped the dynamometer with palm down and steadied it by holding the spring case between thumb and fingers of his other hand. This two-handed support usually increased the score two or three kilograms over that of a one-handed grip, but added to the consistency of performance. Either using talc, or wiping the hand on a towel between pulls, served to keep the palm dry during the test.

The observer first described and demonstrated the procedure to be followed. *S* then adjusted the dynamometer to the preferred size for his hand and assumed a position of readiness to pull. The observer next said, "I will call numbers in order beginning at nine and going up. I will call a number every three seconds. Each time I call a number you must pull to the scale marker of that number. As soon as you have pulled up to it, release your grip, wipe your hand if you wish, and prepare for the next pull. Be sure you pull at least as high as the number I call. Squeeze promptly when I call the number. If you go as much as one number higher than I call, push the maximum indicator back to the one I called, before you grip again. This is to make sure you get up to the next number when I do call it. Keep up with the numbers as long as you possibly can. If you fail to reach a number, try once more when the next number is called. If you are still not up to the one I called, then put your dynamometer down to show you are through. You will do this test once with each hand. Use your right hand first."

The observer said, "Ready," as he started a stop watch. At three seconds he said, "Nine"; at six seconds he said, "Ten"; etc. He continued counting until all the *Ss* who were being tested had reached their limits and then he recorded their scores. The procedure for four *Ss* on one hand, including recording, required $1\frac{1}{2}$ to 2 minutes.

The same procedure was repeated immediately with the left hand after any necessary adjustments of the size of the grip. Each score was read to the nearest third of a scale division (to the nearest kilogram). An *S's* score for a test was the average of the scores of his two hands, in kilograms.

In testing women, whose average grip is somewhat less than that of men, the observer began counting, and *S* began pulling, at "Six." Other details of procedure remained the same.

Subjects. In the determination of test reliability, a standardization group of 72 unselected male naval personnel were tested twice, with 6 to 48 hours elapsing between tests. The first test scores of 97 additional male naval personnel are available from data gathered at the Medical Field Research Laboratory, Camp Lejeune, North Carolina (5). Thirty-two of these *Ss* were tested twice. The scores of 161 Waves in the medical department were also obtained. Data on 648 industrial workers in two TNT plants were supplied by Passed Assistant Sanitarian (R) Robert B. Malmo, USPHS, Industrial Hygiene Research Laboratory,

National Institute of Health. The industrial workers were tested on the preferred hand only. Age statistics on the various groups of Ss are included in Table 1.

Results

Reliability. Reliability coefficients (r) in the standardization group ($N = 72$) are as follows:

	First Test	Second Test
(1) Right and left hand	.79	.78
(2) Preferred and non-preferred hand	.83	.85
(3) Mean of the two hands, (2) corrected for double test length	.91	.92
(4) Right hand, retest		.83
(5) Left hand, retest		.83
(6) Mean of two hands, retest		.87

The scores of the 32 Ss of the Camp Lejeune group on whom a second test is available yield a raw retest correlation coefficient of 0.84.

Score Distributions. In Table 1 are distributions of first test scores for the groups tested. The data for the industrial groups are for the preferred hand only. The comparable preferred-hand value for the mean of the standardization group is 57.14 kg. ($\sigma = 7.44$); for the Camp Lejeune group, 55.97 kg. ($\sigma = 6.81$); and for the Waves, 34.86 kg. ($\sigma = 3.07$). These values are significantly larger ($P < .01$) than those of the comparable men's and women's industrial groups.

Because of the differences in the age distribution of the naval and industrial groups (Table 1), and in view of the significant regression of dynamometer score on age (3, 8), a more accurate estimate of the difference between the naval and industrial populations was desirable. Accordingly, two new mean values for the industrial men were calculated, each corrected for age distribution to match one of the samples of naval men. To secure one of these corrected means, the average dynamometer scores of each five-year age range of the industrial group was weighted by the number of men in that age range in one of the naval samples. Both t -values between preferred-hand scores (calculated by comparing each group of naval men with the industrial group, when corrected for age in this fashion) showed a highly significant superiority ($P < .01$) of the naval personnel. The Waves were similarly compared with the industrial women, and the former were also found to be significantly superior ($P < .01$) when matched for age.

Improvement with Practice. Three groups of Ss have been given the hand dynamometer test over a period of days so that improvement with

Table 1
Distribution of Hand Dynamometer Scores

Score Interval (kg.)	Naval Personnel (Mean of two hands)		Industrial Personnel (Preferred hand score)	
	Men	Women	Men	Women
75-77.9			2	
72-74.9	1		4	
69-71.9	3		10	
66-68.9	5		17	
63-65.9	5		32	
60-62.9	17		46	
57-59.9	27		82	
54-56.9	29		102	
51-53.9	27	1	96	
48-50.9	31	1	63	
45-47.9	14	3	52	1
42-44.9	5	5	30	5
39-41.9	4	21	12	15
36-38.9	1	36	3	17
33-35.9		41	1	30
30-32.9		40		16
27-29.9		11		5
24-26.9		1		6
21-23.9		1		1
<i>N</i>	169	161	552	96
\bar{X}^*	54.0	33.1	53.1	33.6
Mdn.*	54.0	35.0	49.3	35.0
σ^*	6.5	4.0	7.0	4.7
Age of subjects				
\bar{X}	21.8	22.7	34.4	32.4
Mdn.	20.5	22.1	33.5	31.4
σ	4.6	3.3	8.7	8.4
Range	17-46	20-36	18-68	20-57

* These statistics were calculated from the raw scores and will not agree exactly with those derived from the above distributions.

practice can be followed (Fig. 1). Data of two of these groups (curves A and C in Fig. 1) represent performance during a period of training prior to the beginning of an experiment. Data for the other group (curve B) are combined from an experimental and a control group, the former of whom underwent an experimental regime which did not produce significant performance differences from the latter. There is considerable agreement among the three curves as to slope during the first dozen trials in spite of differences in frequency of testing. The average improvement of 40 Ss between the second and twelfth test periods is 0.31

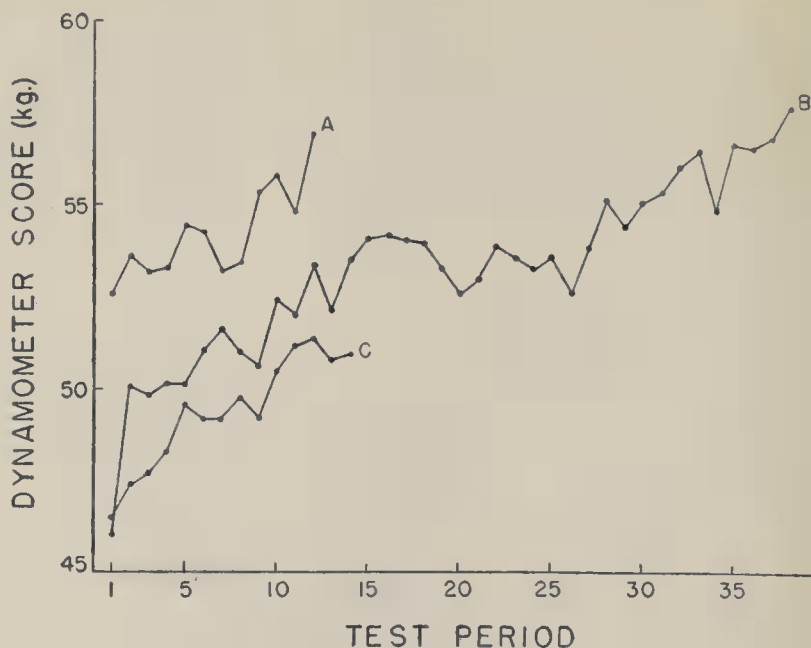


FIG. 1. Hand dynamometer practice curves. Curve A: three tests per day, 6 hours apart, for four days; 18 Ss. Curve B: one test per day for 45 days, except a 6-day interval between tests 1 and 2, and a 3-day interval between tests 2 and 3; 12 Ss. Curve C: one or two tests per day in a period of 17 days (no test on some days); 10 Ss.

kilograms per period. This number is the slope of the line fitted to these eleven points by the method of least squares.

Discussion

The procedure for use of the hand dynamometer developed in this study has shown itself, from the point of view of both split-half and retest reliability, to be sufficiently precise for use with small groups of Ss. The higher reliability of this procedure as compared with others using the hand dynamometer is presumed to result from: (1) the use of a number of trials of increasing difficulty so that the S has opportunity to "warm up," get used to the task, and still have several opportunities to exert himself to the limit; (2) the small amount of fatigue which working near the limit for several trials produces, and which therefore, requires the final squeeze to be made under a standardized stress; and (3) the nearly complete elimination of pain as a factor in determining the score. No special significance is attached to the exact time interval (3 sec.) or the increment (3 kg.) between squeezes in the procedure developed.

The phenomena associated with practice on the test are those to be expected with repeated brief exercise of any muscle or muscle group. In the early part of a practice series, minor changes in the manner of pulling

may be of some significance in score changes, but these are uncommon. The relatively large change between first and second test in one experimental group (curve B, Fig. 1) is probably to be accounted for by the fact that these Ss took their first test under somewhat unfavorable physiological circumstances but had five days of rigidly controlled living and physical exercise before the second.

Applications of the Test

The dynamometer test has "face validity" but its inclusion in a battery of performance tests should nevertheless be justified by some more specific idea of what it measures. Observations made on subjects' behavior in more than 2,000 tests with the hand dynamometer at the Naval Medical Research Institute, under many conditions, are in agreement with factorial findings (2, 4, 9) that there is more than simple strength in such a test and that motivational and attitudinal elements usually play a part in determining a score. More objective indication of the validity of the test as a measure of response to severe physiological stress is available in the data of two experiments conducted at the Institute, and another, using the same procedure, carried on elsewhere.

A group of 105 men were tested on the hand dynamometer before engaging in a "fatigue run" lasting 18 hours. This consisted of nearly continuous marching, calisthenics, and active military exercises, and involved a minimum of hand work. About half the men were forced to drop out before the end of the run on account of fatigue. Thirty-two of these men were retested on the dynamometer one to three hours afterward and before they had any sleep.

The mean decrease in score of this group was 2.03 kilograms, with a standard error of 0.25 kilograms. This decrease in score is to be compared with a mean increase from first to second test of 0.88 kilograms (standard error, 0.40 kg.) in the standardization group. The difference between these two changes is highly significant statistically. Only 38 per cent of the standardization group did not show improvement between the first and second test, whereas 75 per cent of those who failed to finish the fatigue run showed no improvement. In this experiment the physiologically determined ability to perform work was presumably adversely influenced and motivation either was not sufficient to overcome the physiological decrement or was concomitantly decreased.

In a second experiment, 17 Ss were in a closed chamber for two and a half days, during which time the carbon dioxide concentration built up to five per cent and the oxygen decreased from 21 to 12 per cent (1). During this time the Ss were tested repeatedly on a number of "mental" and relatively complex sensorimotor functions which showed only small

or insignificant decrements. The hand dynamometer, however, showed a decrease in score which was highly significant statistically. The mean score just before release from the stress was 1.06 σ below the mean of the pre-experimental distribution and only one *S* showed an increase in score at that time over his pre-experimental value. In this experiment, concomitant measures of cardiovascular and respiratory functions showed several rather severe disturbances. These dynamometer data are interpreted to mean that the *Ss*' "willingness to exert effort" (2, 9) was unable to compensate for the physiological decrement suffered. That such "willingness" was present was inferred from intimate observations of the *Ss* and from the fact that on other tests where motivation was important and strength was not, the *Ss* maintained or nearly maintained their previous performance levels. Mean hand dynamometer performance was still 0.56 σ below the mean of the pre-experimental test four to six hours after release from the chamber and only 5 of the 17 *Ss* had exceeded their pre-experimental performance. A number of the respiratory and cardiovascular functions measured at the same time also failed to recover.

In a third experiment (6) environmental temperature was the variable and was carefully controlled for a period of 44 days. There was no significant difference in daily tests between an experimental and a control group on a wide variety of sensorimotor tasks, and none but minor and easily reversible differences on a number of physiological measures. The hand dynamometer also showed no differences between the experimental and control groups. The curves for the two groups were essentially parallel throughout the experiment and are combined in Figure 1, curve B. The departures from a smooth curve are most simply and adequately explained as motivation changes; a lapse in interest in the middle of the experiment produced the long plateau, with revived enthusiasm being reflected in improved performance as the end of a long confinement approached. These trends were shown to be unrelated to any physiological changes during the same periods.

The first two of these three experiments indicate that this test can measure changes in hand strength when the internal physiological environment is modified beyond the limits for which motivation can compensate. The third experiment indicates the converse: when the physiological state is stabilized over a considerable period, motivation becomes paramount in determining day-to-day variations in individual and group scores.

Summary

1. A test procedure for the hand dynamometer was developed which is satisfactory for inclusion in a battery of performance tests. In this

procedure, *S* squeezes the dynamometer every three seconds, starting with a grip of 27 kg. and increasing his grip 3 kg. with each subsequent attempt. The test continues until he is unable to achieve a level required, the score being the highest level reached.

2. The split-half reliability coefficients for this test procedure, preferred vs. non-preferred hand, were 0.01 and 0.92 for the first and second tests, respectively ($N = 72$). The retest reliability of the test, administered twice within two days, was 0.87 ($N = 72$).

3. Improvement with practice on this hand dynamometer test has been followed in three groups for 12, 14, and 38 practice periods under relatively normal conditions. Mean improvement in the early periods of practice for 40 *Ss* was 0.31 kg. per period.

4. Several groups of data involving various stress conditions indicate that the test has some degree of validity, in that test scores parallel the cardiovascular and respiratory responses and reported general fatigue of *Ss* under such conditions.

Received September 24, 1945.

References

1. Consolazio, W. V., Fisher, M. B., Pace, N., Pecora, L. J., Pitts, G. C., and Behnke, A. R. The effects on man of prolonged exposure to increased carbon dioxide and decreased oxygen pressures. Naval Medical Research Institute, Research Project X-349, Revised Report, 1945.
2. Dunlap, J. W. Tests of the "ability to take it." Civil Aeronautics Administration, Division of Research, Report No. 11, Washington, D. C., 1943 (restricted).
3. Fisher, M. B., and Birren, J. E. Age and hand strength (in preparation).
4. Howell, T. H. An experimental study of persistence. *J. abnorm. soc. Psychol.*, 1933, 28, 14-29.
5. New, W. N., et al. Validation of physical fitness tests by evaluation of performance of subjects. Medical Field Research Laboratory, Camp Lejeune, N. C., Research Project X-526.
6. Pace, N., Fisher, M. B., Birren, J. E., Pitts, G. C., White, W. A., Consolazio, W. V., and Pecora, L. J. A comparative study of the effect on men of continuous versus intermittent exposure to a tropical environment. Naval Medical Research Institute, Research Project X-205, Report 2, 1945.
7. Quetelet, A. *Sur l'homme et le développement de ses facultés*. Paris: Bachelier, Imprimeur-Libraire, 1835, 2 vols.
8. Todd, T. W. Skeleton and locomotor system. In: E. V. Cowdry (Ed.). *Problems of ageing* (2nd Ed.). Baltimore: Williams and Wilkins, 1942 (Chapter 12).
9. Wherry, R. J. Preliminary report on the construction of a test battery of persistence. Supplement I in (1) (restricted).
10. Whipple, G. M. *Manual of mental and physical tests. Part I: Simpler processes* (2nd ed.). Baltimore: Warwick and York, 1914, pp. 100-129.

The Time Appreciation Test *

John N. Buck

Chief Psychologist, Lynchburg State Colony, Virginia

As Dr. Grace H. Kent ¹ has stated, one of the principal purposes of an "emergency test" is to provide the psychiatrist with a psychometric tool that he can employ when no qualified psychological assistance is available.

There are many situations, however, in which such tests can be of very real service to the psychologist himself. Wherever and whenever "time is of the essence," so to speak, the "emergency test" has a definite place. In the clinic and in the mental hospital it may be used as an initial screening device and to help to determine what further psychometric procedure may be indicated; it may also be used at regular intervals as a follow-up check upon the patient's progress. In personnel work it may be employed by the interviewer as an aid in making a rough appraisal of the applicant's qualifications for a given position; also to check upon the veracity of the applicant's statements concerning his educational attainments. The value of the emergency test to the armed forces has been amply attested to in the journals in the last few years.

Standardization

The Time Appreciation Test was first given (as a group test) to approximately 675 white persons ranging in life age from 8 to 23 years, and in educational status from the third grade in Grammar School to the third year in college, in the city schools of Lynchburg, Va.; the Hughes

* The author wishes to acknowledge his especial indebtedness to the following whose fine courtesy and wholehearted cooperation made possible the establishment of the norms: Mr. Omer Carmichael, Superintendent of City Schools, Lynchburg, Va.; Mr. E. W. Paylor, Superintendent, Hughes Memorial School, Danville, Va.; Captain Harry Carmine, Fork Union Military Academy, Fork Union, Va.; Dr. Oscar DeWolf Randolph, Rector, Virginia Episcopal School, Lynchburg, Va.; Dr. William Hinton, Assistant Professor of Psychology, Washington and Lee University, Lexington, Va.; Mrs. Dorota Rymarkiewiczowa, Chief Psychologist, University of Virginia Hospital, Charlottesville, Va. To Mrs. D. E. Mack go deep thanks for her valuable assistance in the statistical treatment of the data. Space does not permit adequate expression of gratitude for the help afforded by many others.

¹ Kent, Grace H., *Oral test for emergency use in clinics*; Mental Measurement Monographs, Baltimore: Williams & Wilkins; 1932.

Memorial School, Danville, Va.; Fork Union Military Academy, Fork Union, Va.; Virginia Episcopal School, Lynchburg, Va.; and Washington and Lee University, Lexington, Va.

The tentative norms thus established on the basis of chronological selection were later amended in certain instances as the result of evidence produced by administering the test individually to some 350 white persons (ranging in life age from 7 to 75 years, and in educational status from "no schooling" to eight years of college work) whose intellectual level had previously been appraised by use of the Stanford-Binet, the Wechsler-Bellevue Scale, or some other psychometric device of comparable status.

In passing, it should be stated that no significant sex differences were found.

The user of the Time Appreciation Test must always bear in mind the fact that the accuracy of the norms has not been established for the negro or the foreign born.

Description

The test consists of 30 questions relating to various aspects of time: the first seven questions deal with immediate orientation; questions No. 17, 18 and 19 are about holidays; questions No. 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, and 30 call for definition of certain phases of time; the remaining nine questions deal with the reduction of large time units into smaller units (and question No. 16 usually demands mathematical calculation of a more than elementary type in the individual test situation).

Tabulation shows that the "rank" order of difficulty of the questions for the Standardization Group was as follows (from the easiest to the most difficult): 1, 8, 3, 4, 6, 9, 12, 10, 13, 14, 2, 5, 18, 7, 23, 21, 15, 25, 17, 16, 19, 26, 20, 22, 28, 11, 27, 24, 29 and 30.

Administration and Scoring

The Time Appreciation Test ² is intended for use with adults ³ and with children of not under ten years life age. It may be used as an individual test, or as a group test (with subjects who have had at least a fourth grade education). All other things being equal, the test is more useful and more accurate when given individually, for the examiner then has an opportunity to obtain information from the subject's response-pattern (decisive or indecisive; rapid or slow; labile or calm, etc.) and verbal-expression (pedantic or colloquial; verbose or terse, etc.) that is

² If the Time Appreciation Test is used as one of a battery of tests, it is suggested that it should not be the last test given.

³ Persons of life age sixteen years or more.

denied him in group examination, and the examiner can clear up, by additional questioning, any ambiguity in the subject's answers.

JNB TIME TEST

(Copyright, 1943, John N. Buck)

Name:.....	Occupation:.....
Sex:..... Race:.....	Residence:.....
Birthdate:.....	Case No.:.....
Education:.....	Examiner:.....

1. Is it morning or afternoon now?
 2. About what time is it by the clock now?
 3. What day of the week is it?
 4. What month is it?
 5. What day of the month is it?
 6. What year is it?
 7. What season of the year is it?
 8. How many days are there in a week?
 9. How many minutes are there in an hour?
 10. How many hours are there in a day?
 11. How many days are there in a month?
 12. How many months are there in a year?
 13. How many seasons are there in a year?
 14. How many seconds are there in a minute?
 15. How many months are there in a season?
 16. How many seconds are there in an hour?
 17. In what month is Thanksgiving Day? On what day in that month does it always come?
 18. In what month is Christmas? On what day in that month does it always come?
 19. In what month is Hallowe'en? On what day in that month does it always come?
 20. What is a decade?
 21. What is a century?
 22. What is a fortnight?
 23. What does anyone mean when he says: "Nine A.M." and "Nine P.M."?
 24. What words do those initials "A.M." and "P.M." stand for?
 25. What does anyone mean when he says: "The year 450 B.C." and "The year 450 A.D."?
 26. What words do those initials "B.C." and "A.D." stand for?
 27. What is a time zone?
 28. Name the time zones in the United States.
 29. What is Greenwich mean time?
 30. What does anyone mean when he says "Vernal Equinox" and "Autumnal Equinox"?
- Score: Correct:.....Half-correct:.....Total Points:.....C.A.....Equiv. M.A.....

FIG. 1. Sample of time appreciation test.

When the test is given to but one subject at a time, the examiner retains the form sheet; reads the questions to the subject, and records the subject's answers verbatim. When the test is given to a number of persons simultaneously, each subject has his own form sheet; reads the questions to himself; and writes his own answers (to the right of the printed questions). See Figure 1.

No explanation or alteration of the wording of the questions (except as indicated later), or of their sequence is permissible. The entire thirty questions should be asked.

There is no time limit; usually the test can be given and scored in ten minutes⁴ or less.

Under the scoring system, as outlined in the following pages, two points are given for each correct answer; and in a number of instances one point is allowed for an answer that is partially correct.

After he has made certain that neither a calendar, nor a watch or clock is available for inspection by the subject, the examiner will say: "Mr. X, I have some questions here that I'd like to ask you. First I am going to ask you a very easy one. . . ."

As a rule no other introductory remarks will be needed, since it is assumed that rapport will have been established before the test is used.

Directions for Scoring

Quest. 1. *Is it morning or afternoon now?* Score two points for correct answer. No partial credit is allowed.

Quest. 2. *About what time is it by the clock now?* Score two points for any answer that is within 30 minutes of the actual time (either way). Example: "1:46" at 1:25. Score one point for any answer that is within 60 minutes (either way) of the actual time. For instance: "2:10" or "12:35" at 1:25.

Quest. 3. *What day of the week is it?* Score two points for the correct answer. In group testing the *number* (as: "second," for Monday) will suffice, but in the individual test situation, the day must be *named* correctly. Score one point for an answer that is not more than one day wrong. For example: on a Tuesday, either "Wednesday" or "Monday" would merit one point's credit.

Quest. 4. *What month is it?* Score two points for the correct answer. In group testing, it will suffice for the subject to give the number (as: "Seventh," for July), but in individual testing the examiner will allow full credit only for the proper *name*, in response to a repetition of the original question. Score one point for the month just ended (on the first day of a new month) as: "July" on the first day of August, or for the month to come (on the last day of a month) as: "August" on the 31st of July.

Quest. 5. *What day of the month is it?* Score two points for the correct answer. Score one point for an answer that is not more than three days wrong (either way). For example: the subject who states on the 31st of March that

⁴ Time consumption greatly in excess of this will usually be significant of a disturbance involving more than a deficiency of intellect.

it is the 28th, 29th, 30th of March or the 1st, 2nd, or 3rd of April will receive a point. Note, though, that he will *not* receive credit at this point if he says it is the 3rd of *March* (not April).

Quest. 6. *What year is it?* Score two points for the correct year. Score one point for the year just ended (on the 1st of January) or the year to come (on the 31st of December). Note: If the examiner can obtain a reply to this question only by adding "It is 1900 and what?", one point is to be given if the subject responds to this "leading" question with the correct year.

Quest. 7. *What season of the year is it?* Score two points for the correct season, or for either winter or spring on the 21st of March; for either spring or summer on the 21st of June; for either summer or fall on the 22nd of September; for either fall or winter on the 21st of December. Score one point for the season to come in the month of the change but *before* the change has actually taken place (as: "Winter" on the 5th of December) or for the season just ended, provided not more than *three* days have elapsed after the change (as: "Fall" on the 23rd of December). Note: If the subject states, for instance: "It is the hunting season," the examiner will continue with: "Yes, I know that, but what season of the *year* is it? Is it so-and-so or so-and-so?" (for "so-and-so" the examiner will substitute the names of the seasons bracketing the actual season—in winter, to illustrate, it would be "Is it fall or spring?")—and if the subject gives the correct season he is to be given one point's credit.

Quest. 8. *How many days are there in a week?* Score two points for either "Seven" or "Six days and Sunday." If the subject answers "Six" or "Six days" or "Six weekdays," the examiner will continue with: "Does that include Sunday" and give the subject one point if he replies that it does *not*. If the subject says, "I don't know," the examiner will ask the subject to name the days of the week. If the subject can do so, the examiner will repeat the original question and if the subject says, "Seven," the examiner will allow one point's credit therefor.

Quest. 9. *How many minutes are there in an hour?* Score two points for "Sixty." No partial credit is allowed.

Quest. 10. *How many hours are there in a day?* Score two points for "Twenty-four" or "Twelve hours in a day; twelve in a night"; score one point for "Twelve." Note: If the subject says, for instance, "There are eight hours in a working day" (or something similar), the examiner will continue with: "But how many hours are there in a full day?" and the examiner will allow one point for "Twenty-four."

Quest. 11. *How many days are there in a month?* Score two points for "28, 29, 30, and 31," or "Anywhere from 28 to 31" (or an equivalent response). Score one point for any answer containing a 30 *and* a 31, and a 28 *or* a 29, but not all four. Note: If the subject says, "Usually 30," or "It varies," or "They have a different number," the examiner will say "Explain," and will allow credit in accordance with the calibre of the reply.

Quest. 12. *How many months are there in a year?* Score two points for "Twelve." No partial credit is allowed.

Quest. 13. *How many seasons are there in a year?* Score two points for "Four." No partial credit is allowed.

Quest. 14. *How many seconds are there in a minute?* Score two points for "Sixty." No partial credit is allowed.

Quest. 15. *How many months are there in a season?* Score two points for "Three." No partial credit is allowed. If the subject says, "That varies with the part of the world you are in," the examiner will ask, "How many months are there in a calendar season?", and he will allow two points credit for the answer "Three."

Quest. 16. *How many seconds are there in an hour?* Score two points for "Three thousand, six hundred," or "Thirty-six hundred." This answer must

be arrived at without the aid of pencil and paper (or their equivalent) when the test is given individually; in group testing the subject may work out the problem on his Form Sheet—in a surprising number of instances this computation is incorrectly done! No partial credit is allowed.

Quest. 17. *In what month is Thanksgiving day? On what day in that month does it always come?* Score two points for "In November; on the last Thursday," or "November, on the Fourth Thursday." If the subject says "In November—on the *third* Thursday," the examiner will say, "Yes, President Roosevelt did proclaim the third Thursday as Thanksgiving, but on what day in the month did the old-fashioned Thanksgiving always come?" and he will allow full credit for "Last," or "Fourth." Score one point if November is named correctly, but the day of the month is not. No credit is to be allowed if the *month* is named incorrectly even though the day itself is given as the last Thursday.

Quest. 18. *In what month is Christmas? On what day in that month does it always come?* Score two points for "December 25th" (or its equivalent). Score one point if the month is named correctly, but the day is not. No credit is to be given if the *month* is not named correctly.

Quest. 19. *In what month is Hallowe'en? On what day in that month does it always come?* Score two points for "The last day of October," "October 31st" (or their equivalent). Score one point if the month is named correctly, regardless of what day is given. No credit is to be granted if any month but October is named.

Quest. 20. *What is a decade?* Score two points for "Ten years," or "Ten days." No partial credit is allowed.

Quest. 21. *What is a century?* Score two points for any answer indicating that a century is a period of 100 years. No partial credit is allowed.

Quest. 22. *What is a fortnight?* Score two points for "Two weeks," or "Fourteen days," or "Fourteen nights." Score one point for "Fifteen days" or "Half a month."

Quest. 23. *What does anyone mean when he says: "Nine A.M." and "Nine P.M."?* Score two points for any answer indicating that the former precedes noon; the latter follows it. The answers: "Afternoon," "Evening," or "Night," all suffice for the latter. Score one point if either "Nine A.M." or "Nine P.M." is defined correctly, but not both.

Quest. 24. *What words do those initials "A.M." and "P.M." stand for?* Score two points for "Ante Meridian and Post Meridian." Note: Misspelling of the words in group testing does not rob the subject of full credit if the examiner feels certain that the fault lies in the spelling alone. Note: If the subject merely repeats his "Morning and night" response to Question 23, and there is any reason to believe he actually knows the correct answer, the examiner should repeat Question 24 casually. Score one point for *either* "Ante Meridian" or "Post Meridian," but not both; or for *both* "Before noon" and "After noon." No credit is to be given for "ante morning" and "post morning."

Quest. 25. *What does anyone mean when he says: "The year 450 B.C." and "The year 450 A.D."?* Score two points for any answer indicating that the former date was 450 years before the birth of Christ, the latter 450 years after Christ's birth. Note: In the individual test, if the subject states that the two terms mean, respectively, "Before and after Christ," full credit cannot be allowed until and unless additional questioning by "What do you mean by *after*? After birth or after death?" elicits the fact that the subject means *after birth*. But in group testing "Before and after Christ" receives full credit since further questioning is impossible. Score one point if either term is properly explained, but not both.

Quest. 26. *What words do those initials "B.C." and "A.D." stand for?* Score two points for "Before Christ" and either "Anno Domini" or "In the

year of our Lord." Note: Misspelling of the words "Anno Domini" does not cost the subject full credit (in group testing) if his intent is obvious. Score one point if either set (but not both) is defined correctly.

Quest. 27. *What is a time zone?* Score two points for an answer stating, in effect, that the world is divided arbitrarily into 24 longitudinal belts of approximately equal size (15° each), with the time in adjacent belts one hour apart. Score one point for an answer that defines a time zone as an area of the earth's surface in all parts of which the time is the same simultaneously. Note: the examiner must be careful to appraise the calibre of the answers on the basis of the subject's knowledge rather than on his ability to express himself concisely.

Quest. 28. *Name the time zones in the United States.* Score two points for: "Eastern, Central, (Rocky) Mountain, and Pacific"—named in any order. Score one point for three of the four zones named above (one of the three *must* be the zone in which the subject is at the moment); or for an answer naming the correct four zones and also a fifth and incorrect zone (such as Midwestern, or Western). The inclusion or omission of "Standard," "Daylight Saving," "War Time," as elaboration (as, for example, "Eastern War Time, Central War Time," etc.) does not affect the scoring.

Quest. 29. *What is Greenwich mean time?* Score two points for any answer that says, in effect, that all standard times are based on their longitudinal relationship to the meridian on which is located the Greenwich Astronomical Observatory in England. Score one point for "That's the time at Greenwich, England" (Greenwich Village will not suffice), or "Greenwich is located at 0° longitude"—or something equivalent thereto. The differential credit factor in this question and in question 27 is the all-inclusiveness or localization of the subject's answer.

Quest. 30. *What does anyone mean when he says: "Vernal Equinox" and "Autumnal Equinox"?* Score two points for any answer that accurately defines "vernal," "autumnal," and "equinox." It is not necessary that the subject be specific as to dates; neither is it necessary for him to define the sun's position in relation to the equator. Score one point for any answer correctly defining "vernal" and "autumnal," or "equinox," but not all three.

As soon as the credit points have been summed up, the examiner can turn to the tentative norms in Table 1 and convert the point score into a mental age equivalent score and if the patient is 16 years of age or older, to an adult intelligence quotient, and the equivalent adult intelligence level classification as well.

Validity

Tables 2 and 3 offer evidence as to the validity of the Time Appreciation Test. The test's comparatively high degree of correlation with the Stanford-Binet seems quite satisfactory for an emergency test.

Reliability

It is believed that the difference between the reliability coefficients shown in Table 4 is due more to the difference between the types of subjects who made up the groups than to the difference between the methods of administration of the test. It is scarcely to be expected, however, that the test-retest reliability coefficient will be as high for group administra-

Table 1

Tentative Norms for the Time Appreciation Test

Note: All four columns are for use with subjects of life age sixteen or above. The first two columns (reading from left to right) may be used with subjects of life age ten and over for converting point scores to mental age equivalent scores.

Time Appreciation Test Point Score	M.A.	Adult I.Q.*	Adult Classi- fication
2	4:6	30	Imbecile
3	5:0	33	
4	5:6	37	
5	6:0	40	
6	6:6	43	
7	7:0	47	Moron
12	7:6	50	
16	8:0	53	
21	8:6	57	
25	9:0	60	
28	9:6	63	Borderline
31	10:0	67	
33	10:6	70	
35	11:0	73	
37	11:6	77	
39	12:0	80	Dull average
41	12:6	83	
42	13:0	87	
43	13:6	90	
44	14:0	93	
45	14:6	97	Average
46	15:0	100	
47	15:6	103	
48	16:0	107	
49	16:6	110	
50	17:0	113	Above average
51	17:6	117	
52	18:0	120	Superior
53	18:6	123	
54 and up	above 18:6	above 123	

* Estimated according to the method of Terman and Merrill, *Measuring intelligence*; Boston: Houghton Mifflin Company, 1937. When comparing the Time Appreciation Test I.Q. of a subject of life age 30 and above with his Wechsler-Bellevue I.Q., it is suggested that the examiner in computing the Time Test I.Q. allow for the age factor by using the "Table of Approximate C.A. (Adult M.A.) Denominators for Binet Scales . . ." as described in the third edition of *The measurement of adult intelligence*, by David Wechsler (Baltimore: Williams and Wilkins, 1944).

Table 2

Pearson Coefficients of Correlation between the Time Appreciation Test and Three Other Intelligence Tests

Note: Group A was composed of students (9th through 12th grades) of the Virginia Episcopal School of Lynchburg, Virginia. Both the Otis and the Time Appreciation Test were given as group tests. Groups B, C, D, E, and F were composed of in- and out-patients of the Lynchburg State Colony, Colony, Virginia. The patients in question were mentally deficient and/or epileptic, or suspected of being either or both. The tests in each instance were given individually. Group G was composed of mentally deficient in- and out-patients of the Lynchburg State Colony, Colony, Virginia (there were no epileptics in this group). In each instance the tests were given individually.

Group	Chron. Age	Test	No. Cases	r	P.E. _r
A	12 to 19	Otis, Gamma, Form A, I.Q.	103	.67	$\pm .037$
B	16 and over	Wechsler-Bellevue Total I.Q.	59	.74	$\pm .041$
C	16 and over	Wechsler-Bellevue Verbal I.Q.	22	.67	$\pm .080$
D	16 and over	Wechsler-Bellevue Performance I.Q.	24	.77	$\pm .056$
E	16 and over	Stanford-Binet (Form M) I.Q.	32	.88	$\pm .027$
F	Below 16	Stanford-Binet (Form M) I.Q.	36	.78	$\pm .044$
G	Below 16	Stanford-Binet (Form M) I.Q.	31	.84	$\pm .035$

tion as for individual administration unless, somehow, the subjects to whom the test is given as a group can be prevented from "comparing notes" afterwards.

Table 3

Comparison of Measures of Central Tendency and Variability

Group	Chron. Age	Tests	No. Cases	Mean	Range	σ
A	12 to 19	Otis, Gamma, Form A, I.Q.	103	103	82-124	12.90
		Time Appreciation Test I.Q.		98	75-123+	9.00
B	16 and over	Wechsler-Bellevue Total I.Q.	59	61	36-90	12.37
		Time Appreciation Test I.Q.		61	37-83	9.15
C	16 and over	Wechsler-Bellevue Verbal I.Q.	22	67	52-79	8.05
		Time Appreciation Test I.Q.		61	40-80	9.34
D	16 and over	Wechsler-Bellevue Performance I.Q.	24	59	34-94	14.00
		Time Appreciation Test I.Q.		59	47-80	8.00
E	16 and over	Stanford-Binet (Form M) I.Q.	32	48	22-73	13.98
		Time Appreciation Test I.Q.		53	26-80	13.60
F	Below 16	Stanford-Binet (Form M) I.Q.	36	48	26-68	11.80
		Time Appreciation Test I.Q.		51	27-76	12.38
G	Below 16	Stanford Binet (Form M) I.Q.	31	47	26-68	11.70
		Time Appreciation Test I.Q.		50	27-76	13.40

Table 4

Test-Retest Reliability for the Time Appreciation Test

I. Group Administration of Test

Note: The subjects in this group were 48 student nurses at the University of Virginia Hospital. Seventeen days intervened between test and retest.

	Mean	Range	σ	r	P.E.,
First Test	98	77-120	10.68	.80	$\pm .037$
Second Test	104	80-123	11.08		

II. Individual Administration of Test

Note: The subjects in this group were 20 patients at the University of Virginia Hospital or the Lynchburg State Colony (1 psychoneurotic; 1 paranoid condition; 1 hypoparathyroidism; 4 epileptic; 13 mentally deficient). The test-retest time interval averaged 33 days; ranged from 4 days to 16 weeks.

	Mean	Range	σ	r	P.E.,
First Test	61	30-100	16.09	.96	$\pm .012$
Second Test	64	33-107	18.27		

Interpretation

It is assumed that in the great majority of instances the subject's point score will represent, albeit crudely, the subject's effective level of intelligence at the moment that he takes the test. This level, of course, will not always be his "normal" effective level.

It appears that one may assume that answers that fall into one or more of the following categories are indicative of *potential* ability: (a) answers whose incorrectness appears to be due to lack of attention rather than to any real lack of knowledge (to illustrate: the answer, "365" to question No. 12); (b) mistakes in answer to one or more of the first seven questions where the score on the remaining twenty-three questions is of good calibre; (c) answers to questions No. 20, 21, 22, 27, 28, 29 and 30 that are not of high enough calibre to merit either one or two points but that show some familiarity with the material involved.

The examiner will find it worthwhile to total point scores for the first ten, the second ten, and the third ten questions separately: the point score of the average subject is highest for the first ten, progressively lower through the next two tens. Any change in this order suggests the presence of some disintegrating factor and the need for more careful examination of the subject.

Such frankly bizarre responses as: "Some kind of weeks are good," in reply to question No. 3, or "Same place I live in," for question No. 18, strongly suggest the presence of a psychosis.

Answers that are accompanied by frequent, “. . . , isn't it?”; “. . . could be . . .”; “I think there are . . .”; etc., are usually, as they appear to be, indicative of anxiety.

Experience so far has indicated that the average subject's point score tends to hold up well with advancing life age (though not so well as his vocabulary score, for example). The point score appears to be rather sensitive to psychic disturbances resulting in aprosexia.

It must always be borne in mind that this is an “emergency test”—no more. A competent psychologist should not need to be told that no positive diagnosis may ever be made solely on the basis of the subject's score on this or any other emergency test.

Summary

The Time Appreciation Test is composed of thirty questions all of which touch in some way upon some phase of time.

In its favor appear to be the following points: 1. Ease of administration; 2. Simplicity of scoring; 3. Economy of time; 4. The fact that no special testing equipment is required; 5. Its relative freedom from cultural artifacts (much of the information sought is of a type that seems to be acquired by the average individual in the normal course of growing up)⁵; 6. The questions are apparently inoffensive to the average subject.

Opposed to this are the following: 1. A probable sampling inadequacy in the groups on which the norms were established; 2. Overweighting of several of the last ten questions with some “special knowledge” factors; 3. The fact that several of the questions are “seasonal” in type (that is, their relative difficulty is not uniform throughout the calendar year).

Ultimately it may prove possible to correct certain of these defects.

It is hoped that with all its faults the Time Appreciation Test will prove itself a useful addition to the emergency test group.

Received August 19, 1945.

⁵ Illiterates seem to be penalized somewhat less severely on this than on most verbal tests.

Relation of Iowa Silent Reading Test Scores to Measures of Scholastic Aptitude and Achievement *

Richard W. Kilby

Woman's College of The University of North Carolina

As a further step toward understanding what the Iowa Silent Reading Test measures and its relation to college grades a random sample of 100 Yale freshmen was studied.¹ Correlations were run between the I.S.R. Test (total and subtest scores) and final grades and various aptitude measures. The randomizing procedure consisted of listing the class alphabetically and using every fourth case for the sample.

The correlations between the I.S.R. Test scores and final grades are presented in Table 1, and the correlations between the I.S.R. Test scores and various measures of aptitude and ability are given in Table 4.

Correlations between the I.S.R. Test scores and final grades will be considered first. All subtests except Rate and Paragraph Comprehension are significantly correlated with average final grade. Rate bears a very low and insignificant relation to grade standing; Paragraph Comprehension falls a little short of significance. Imus, Rothney, and Bear (1, pp. 67, 122) found a similar absence of relation between grades and rate of reading as measured by the I.S.R. Test, and concluded that academic performance is seldom improved by increasing the rate of reading. Such a conclusion is not warranted, without qualification, by the low correlation found; before that conclusion should be drawn it would have to be demonstrated that an increase in rate of reading according to several different types of rate measure is not correlated with improvement in grades.

In view of the low correlation of rate of reading with final grades an investigation of the validity of the I.S.R. Rate subtest seemed advisable. This was made possible by studying a different group of students who had participated in a remedial reading program and had used homogeneous practice materials (Equated Transfer Selections, distributed by the Harvard Film Service, Biological Laboratories, Harvard University)

* This paper is part of a dissertation which was presented for the degree of Doctor of Philosophy in Yale University. I am indebted to J. Richard Wittenborn, Clinical Psychologist, for permission to use the research facilities of the Division of Student Mental Hygiene.

¹ Research upon the Iowa Silent Reading Test has recently been reviewed by Triggs (3).

for speed reading during practice sessions and had been given a final I.S.R. Test to measure improvement. Since the students had recorded on progress sheets their reading rates for the practice material, it was possible to determine the reliability of the practice materials and use them for validating the I.S.R. Rate subtest. Fifty-nine cases were involved. The average of the last six practice selections read before taking the test was used in the correlation. The reliability coefficient of

Table 1
Correlations between Iowa Silent Reading Test Scores and Final Grades**

Grade	N	I.S.R. Subtests								Median Standard Score
		1		2	3	4	5	6	7	
		a-c	b-d							
Ave. of finals	100	.03	.24*	.45*	.30*	.42*	.40*	.22	.31*	.40*
English	97	.03	.12	.31*	.24*	.38*	.39*	.26*	.24*	.31*
Social sciences	48	.17	.32	.43*	.32	.56*	.57*	.38*	.26	.57*
Sciences	85	.00	.19	.28*	.21	.22	.29*	.12	.28*	.26
Languages	64	.08	.09	.28	.34*	.39*	.35*	.13	.22	.25
Mathematics	83	.06	.19	.33*	.14	.25	.23	.13	.24	.26

* Asterisks indicate those coefficients that are at least four times their probable errors.

** The probable error of any of the correlations in the above table may be readily determined from the following table:

Range of r 's having the respective P.E.'s:						
N	.09	.08	.07	.06	.05	.04
48	.00-.35	.36-.47	.48-.57	.58-.65		
64		.00-.32	.33-.47	.48-.58		
83			.00-.34	.35-.50	.51-.62	
85			.00-.32	.33-.49	.50-.61	
86			.00-.32	.33-.49	.50-.61	
97			.00-.22	.23-.44	.45-.58	.59-.69
99			.00-.20	.21-.43	.44-.57	.58-.69
100			.00-.18	.19-.42	.43-.57	.58-.69

the practice material (the last three odd selections against the last three even selections) was .90. The coefficient of correlation between rate on the practice material and rate on the I.S.R. Rate subtest was .69 (P.E. — .05), and when corrected for attenuation became .75. This correlation indicates a reasonably high validity for the I.S.R. Rate subtest, but it should be pointed out that the validating selections were of approximately the same level of difficulty as the I.S.R. Rate subtest, and rate of reading was measured in the same way, i.e. number of words read each minute.

It would be safest to interpret the correlation as indicating that the rate subtest has reasonably high validity for measuring that type of rate of reading.

The correlation of grades in various types of courses with Rate subtest are uniformly low, indicating that a student's rate of reading as measured by this test bears no relation to his success in any type of subject. On the other hand, subtests 2 (Directed Reading), 4 (Word Meaning), and 5 (Sentence Meaning) tend to be uniformly related to grades in all courses. Various other subtests are significantly related to grades in one type of course but not another.

Some light is cast on the validity of the I.S.R. Test by the correlations of the various grades with the total reading score; the degrees of relationship shown are as would be expected for a test whose purpose is to meas-

Table 2

Zero Order and Partial Correlations between the Iowa Silent Reading Test, Scholastic Aptitude Test, and Average Final Grade

1. Iowa Silent Reading Test											3.
N = 100	1									S.A.T.	
	a-c	b-d	2	3	4	5	6	7	Md.		
2. Ave. final grade	.03	.24	.45	.30	.42	.40	.22	.31	.40	.44	
3. S.A.T.	.07	.14	.37	.33	.42	.60	.24	.12	.43		
$r_{12.3}$.00	.20	.36	.19	.30	.20	.14	.30	.27		

ure general reading ability. At the same time, the correlations show that this test measures in large part those reading abilities needed for courses in English and the social sciences rather than in the physical sciences, mathematics, or languages.

The possibility that all subtests other than the Rate subtest were measuring a verbal factor rather than reading ability and that this factor accounted for the relation to final grades made it advisable to find the partial correlations between the I.S.R. Test and grades when scholastic aptitude, as measured by the Scholastic Aptitude Test, is held constant. The partial correlations of the I.S.R. subtests with grades when General Prediction score (predicted grade) is held constant were also determined, since General Prediction score would tend to account for most of the factors related to grades except reading ability. The sets of partial correlations, presented in Tables 2 and 3, indicate that some of the abilities measured by the I.S.R. Test continue to be related to final grades when other factors are partialled out. Regarding first the inter-

Table 3

Zero Order and Partial Correlations between the Iowa Silent Reading Test, General Prediction Score, and Average Final Grade ($N = 100$)

	Iowa Silent Reading Test									
	1		2	3	4	5	6	7	Md.	G.P.
	a-c	b-d								
2. Av. final grade	.03	.24	.45	.30	.42	.40	.22	.31	.40	.69
3. Gen. prediction	.04	.18	.42	.32	.42	.50	.14	.17	.33	
$r_{12.3}$	-.01	.17	.26	.11	.21	.09	.18	.28	.26	

relations between S.A.T., final grade average, and I.S.R. subtests, certain subtests have a higher partial correlation with grades than do others; the rate section of subtest 1 has no relation, while subtests 2 (Directed Reading, 4 (Word Meaning), and 7 (Location of Information) have the highest relation to grades.

Regarding the interrelations between the second group of variables—General Prediction score, final grade average, and I.S.R. subtests—again subtests 2, 4, and 7 have the highest correlation with grades when predicted grade is partialled out, and Rate shows the same absence of relation.

Further knowledge as to what the I.S.R. Test measures may be gained from comparing it with various measures of aptitude or ability. This

Table 4

Correlations between Iowa Silent Reading Test Scores and Measures of Aptitude and Ability**

Measure	N	I.S.R. Subtests								Median Standard Score
		1		2	3	4	5	6	7	
		a-c	b-d							
Gen. prediction	100	.04	.18	.42*	.32*	.42*	.50*	.14	.17	.33*
S.A.T.	100	.07	.14	.37*	.33*	.42*	.60*	.24*	.12	.43*
M.A.T.	100	.00	.18	.18	.14	.24*	.22	.02	.11	.08
Yale Spatial- Mech. Apt.	99	.26*	.26*	.24*	.36*	.25*	.20	.22	.14	.32*
C.E.E.B. Eng- lish Essay	86	.10	.03	.21	.16	.11	.24	.14	.08	.21

* Asterisks indicate those coefficients that are at least four times their probable errors.

** The probable error of any of the correlations in the above table may be readily determined from the table of probable errors in the footnote of Table 1.

was done by correlating it with the College Entrance Examination Board English Essay Examination, the Scholastic Aptitude Test, the Mathematical Aptitude Test, the combined Spatial and Mechanical Aptitude Tests of the Yale Battery, and General Prediction score. The resulting correlations are presented in Table 4. It is evident from these data that this reading test does measure in varying degrees something other than the things which these other tests purport to measure. The correlation of subtest 5 (Sentence Meaning) with S.A.T. is significantly higher than are those of subtests 1 (Rate and Comprehension), 6 (Paragraph Comprehension), and 7 (Location of Information).

While reading tests have been used mainly in entrance examination batteries to locate poor readers, it would be well to know whether or not the same reading test score might also be of value in improving the prediction of college success. At present the average multiple correlation

Table 5

Intercorrelations, Partial, and Multiple *R* between the Iowa Silent Reading Test, Average Final Grade, and General Prediction Score*

<i>N</i> = 100	2.	3.
	I.S.R. Median	Ave. Final Grade
1. General prediction	.33	.69
2. I.S.R. median		.40
$r_{23.1}$.26	
$r_{13.2}$.66	
$r_{3(12)}$.80	

* See footnote, Table 1, for the probable error of any of the above zero order correlations.

coefficient of combinations of tests for predicting college scholarship is about .65, and the range of the middle 50 per cent of coefficients that have been reported is from .60 to .70 (2, 4). Predictive coefficients need to be much higher, but investigators feel that until grades are made more reliable, predictive efficiency cannot increase much if at all. However, if it can be shown that reading ability is correlated with grades when other predictive factors have been partialled out, then it may be said to have value as a predictive addition.

It is unlikely that most reading tests would make a predictive addition, because they duplicate too closely what is already being measured by a scholastic aptitude test, as is indicated by the generally high correlations between reading and aptitude tests. The I.S.R. Test, however, is not so highly correlated with aptitude tests as are other reading tests; for the sample studied here a correlation of only .43 with the Scholastic

Aptitude Test was found (Table 4). This correlation indicates that the I.S.R. Test is measuring something in addition to what it measures in common with the S.A.T. and suggests that the I.S.R. Test may be independently related to grades. To investigate this possibility the intercorrelations, partials, and multiple correlation for the three variables—I.S.R. Test median, General Prediction score and average final grade—were computed. These data are presented in Table 5. The correlation of I.S.R. Test median with final grade average is .40 and the independent relation—the partial correlation—with average final grade when the duplication with General Prediction score has been eliminated is .26. The multiple *R* involving the combination of General Prediction score and I.S.R. Test median is .80, which is high for this type of relationship. These results suggest that the addition of the I.S.R. Test median would be a valuable contribution to a predictive battery since it is independently related to grades. However, these findings must be verified on a much larger population before they may be accepted.

If the I.S.R. Test is found to be of predictive value it is likely that certain subtests, because of higher independent correlations with grades, will be of greater value than others. The data already presented in Tables 2 and 3 suggest that subtests 2, 4, and 7 may be of value because they have the highest partial correlations with grades, while others of the subtests, because of their low partial correlations, are of no predictive value and might well be omitted.

Summary

1. All I.S.R. subtests except the Rate subtest are related to final grades.
2. The Rate subtest possesses reasonably high validity when validated upon material of about the same level of difficulty using the same method of measurement.
3. The correlation of grades in all types of courses with rate of reading as measured by the Rate subtest was uniformly low.
4. The I.S.R. Test was found to have a higher correlation with grades in English and the social sciences than with grades in the physical sciences, mathematics, and languages. The former correlations were not significantly higher than the latter.
5. The I.S.R. Test possesses an independent relation to final grades when other variables are partialled out.
6. The I.S.R. Test measures something other than is measured by various aptitude tests.
7. Use of the I.S.R. Test median in a battery for predicting scholastic success may increase the accuracy of prediction.

8. Certain I.S.R. subtests (2, 4, and 7) are more highly correlated with grades than are others, when related variables are partialled out.

Received September 14, 1945.

References

1. Imus, H. A., J. W. M. Rothney, and R. M. Bear. *An evaluation of visual factors in reading*. Hanover, New Hampshire: Dartmouth College Publications, 1938, pp. 144.
2. Segal, David. Prediction of success in college. *G. P. O. Bull.*, 1934, No. 15, Office of Education.
3. Triggs, Frances O. *Remedial reading*. Minneapolis: University of Minnesota Press, 1943, pp. 219.
4. Wagner, M. E. Prediction of college performance. *University of Buffalo Studies*, 1934, 9, 125-144.

The Relationship of College Board Examination Scores and Reading Scores for College Freshmen

Helen E. Peixotto

Wheaton College, Norton, Massachusetts

This study is an attempt to find a method of preliminary screening of poor readers by means of the College Board Examinations for freshmen entering a liberal arts college. There are various screening tests for reading, but since these tests are usually given by the colleges, the results are not available at the beginning of the first semester. Therefore it would be beneficial to a college remedial reading program to be able to make a preliminary survey of reading ability among the freshmen at the opening of the first semester.

Definition of Terms

The tests used in this study are the College Board Examinations and the Cooperative English Test C2: Reading Comprehension. The meaning of the various scores derived from these tests is important for any interpretation of the data. This is particularly true of the reading scores where the technique of administration influences the meaning of the score, and although many investigators in making tests have used similar terminology the technique of obtaining the scores has varied widely. Discussion of this problem of terminology with reference to speed of reading has recently been reported (1, 3, 4). Therefore the meanings of the scores from the tests used in this investigation are given, and wherever possible, as direct quotations by the authors of the tests.

College Board Examinations:

"The scores on both the achievement tests and the Scholastic Aptitude Test give the schools information valuable to diagnostic and guidance purposes" (5).

In other words the Scholastic Aptitude Test score presumably measures aptitude to do college work. A definite statement regarding this test is difficult to find, but from the reports its validity is based on success of students in college work as reported by various colleges. The definition of its function is further complicated by the fact that the Scholastic Aptitude Test is divided into verbal and mathematical sections, each yielding separate scores, but all studies and reports refer to a Scholastic Aptitude Test score without indicating which Scholastic Aptitude Test

score has been used. The following quotations may give some idea of the ability tapped by the test by means of comparisons—

"The Social Studies Test samples all that a student has learned throughout his schooling and in his outside reading in the social studies. It is a measure of cumulative achievement and growth. It is definitely not a measure of accomplishment in one particular course.

"... the Social Studies Test is suitable for students with different amounts of training so long as the results are interpreted in the light of the amount of preparation. Even without such interpretation, Social Studies Test scores, as we have seen, correlate as highly with freshmen grades as the Scholastic Aptitude Test. If the amount of study is taken into account the test becomes a still better predictor of success in college.

All in all, the evidence from this study points to the fact that the Social Studies Test is a measure of ability and past achievement in the field of social studies, and that it is also a good index of future achievement in College" (7).

For the English Essay Test, too, it is difficult to find any exact statement of just what is being measured. The best description seems to be contained in the following phrase, "accuracy and clarity in writing" (8).

The Cooperative English Test C2: Reading Comprehension is best understood by the following quotations:

"Vocabulary Score indicates the extensiveness of the individual's word knowledge."

"Speed of Comprehension Score represents the product of the rate at which the individual has attempted to comprehend the test material and his success in comprehending it. It is not . . . merely a measure of the number of words read without regard to the thought content."

"Level of Comprehension Score provides a measure of the ability of the student to comprehend materials of increasing difficulty at the rate at which he chooses to work. It is a measure of 'power' or 'depth' of comprehension, indicating the extent to which a pupil is able to grasp the full import of what he reads."

"Total Reading Score is a composite score in which each of the other three scores has equal weight. It may be regarded as a measure of linguistic ability and should prove to be an excellent index of scholastic aptitude" (2).

Procedure

Each freshman in this investigation took College Board Examinations in one of the several centers throughout the country. The College Board Examinations, as stated above, consist of an aptitude examination from which there are two scores, verbal and mathematical. The verbal score is the only one reported in this study. There are also certain achievement tests, but all the students do not take all of these. All the students in this group took the English Essay Test, and their scores are, therefore, utilized in this study.

Some students take these examinations twice, at the end of the junior year in high school and again at the end of the senior year. Others take

them only once, either at the end of the junior or senior year in high school. The reports of the College Board are indefinite as to the relative effects of growth and practice on retest scores, but suggest that perhaps 15 to 20 points should be deducted for practice effects when the tests are taken one year apart (6). Therefore the first score achieved by each student on these tests is used.

During the first week of college each freshman was given the "Cooperative English Test C2: Reading Comprehension." One group of students took Form Q of this test, the other Form R. The two forms are reported as comparable by the Cooperative Test Service. Therefore the results of the two groups are pooled in this paper as if all the students had taken the same Form. Four scores, as described above, are obtained from this test: Vocabulary, Speed of Comprehension, Level of Comprehension and Total Score.

Intercorrelations of the various scores, i.e., verbal Scholastic Aptitude Test, English Essay Test and the reading test, have been computed from the scores of 263 students, members of two classes of freshmen.

Results

The results of this study are presented in terms of intercorrelations for 263 girls in a liberal arts college. This group represents the combined scores of the freshman class for the year 1943 and the year 1944. Since the mean scores and standard deviations for the two groups are closely similar only the composite table is reported here. These intercorrelations are given in Table 1.

Some of the relationships shown in this table are somewhat different

Table 1
Intercorrelations, Means and Standard Deviations of the Six Variables for the
Composite Group ($N = 263$)

	A	B	C	D	E	F
A		.51	.49	.77	.75	.19
B			.74	.86	.60	.16
C				.89	.59	.14
D					.76	.18
E						.21
Mean	67.3	57.8	61.9	63.2	496.0	514.5
Standard Deviation	21.5	25.6	23.8	22.0	74.9	92.7

Legend: A = Vocabulary; B = Speed of comprehension; C = Level of comprehension; D = Total score; E = Verbal Scholastic Aptitude Test; and F = English Essay Test.

from what one might expect. For instance, vocabulary is at least as closely related to speed of comprehension in reading as it is to level of comprehension. It would also appear that vocabulary is an important factor in the Scholastic Aptitude Test since the correlation is so high—as a matter of fact the verbal Scholastic Aptitude Test (the test with which we have chosen to work) is made up of antonyms, analogies and paragraphs (5). It seems apparent, then, that vocabulary is an important factor in aptitude for college work as determined by the Scholastic Aptitude Test.

Level of comprehension and speed of comprehension are closely related, but this is partially due to a spurious factor in the method of obtaining these scores (3). Neither speed of comprehension nor level of comprehension appears to be as important a factor in Scholastic Aptitude Test as is vocabulary.

Since the Total Reading score is a composite of the other three reading scores the correlations between this score and the other three are highly spurious. However level of comprehension seems to be the most important of the three in determining the Total Score. It is evident that Total Scores on the reading test correlate highest with Scholastic Aptitude Test scores, thus substantiating the supposition of the authors of this test, that it “. . . should prove to be an excellent index of scholastic aptitude” (2). It will be noted, however, that this correlation is very little higher than that between vocabulary and the Scholastic Aptitude Test. This raises the question whether the verbal section of the Scholastic Aptitude Test tells us much more than vocabulary would in regard to aptitude for college work.

The English Essay Test correlates with the other tests to a low degree, but all the correlations are significant at the 1% level. Noyes (8) states that he would not expect a high correlation between English Essay Test scores and vocabulary, or Scholastic Aptitude Test scores. His whole discussion seems to be in a hypothetical vein, but he feels that ability in English can be determined with precision when the two scores, English Essay and Scholastic Aptitude, are combined. He suggests no method of weighting, interpretation or procedure in this proposed combination. Apparently one can presume that this test measures a function largely independent of those measured by the other tests.

When the scores of individual students in this study are considered, 65 were in the lowest quartile according to Scholastic Aptitude Test scores, and of these 42 obtained scores below the 35th percentile in one or more of the reading scores. Thus 65% of those in the lowest quartile on the Scholastic Aptitude Test also obtained low percentile ranking on the reading test.

Discussion

There have been other studies which have correlated scores from various tests, but the procedure and purpose of these studies have differed from the present one, although in general the results have been in accord with those found here.

Humber (9) found the relationship between "Honor Point Ratio," general aptitude as measured by the American Council Examination and various reading tests, including the one used in this study, for seniors in various curricula fields. His results show that, with the exception of dietetics, reading scores have greater predictive value for success in college than does general aptitude; reading scores are related to the humanities but infrequently to the sciences. Thus seniors are comparatively homogeneous in aptitude, so that high achievement depends more on reading efficiency than on aptitude as measured by the American Council Examination. Of the correlations with achievement and scores from the Cooperative Reading Test the significant correlations, at the 1% or 5% level, are with Speed of Comprehension, Level of Comprehension and Total Score. There is no significant correlation with Vocabulary which, it will be recalled, was found to correlate higher with verbal Scholastic Aptitude Test scores than did Speed of Comprehension or Level of Comprehension. These findings of Humber seem to corroborate those reported above.

An investigation by Williamson (10) finds a low predictive value for College Aptitude Test in relation to freshman grades and high school scholarship. He suggests possible causes or error which may vitiate the results and offers as a solution increased personnel work to eliminate personal factors among students. One might presume, in view of Humber's study (9) and the results presented here, that reading efficiency might be a significant variable in the relationship studied.

The tests used in this study have been in whole or in part different from those in the two investigations quoted above. Those here studied are tests in national use, which should add to the applicability of the results and, apart from the specific need which this investigation was designed to meet, aid in the interpretation and application of scores derived from them. The results of Williamson (10) and Humber (9) substantiate, though indirectly, the results found here.

Conclusion

From the results of this study it seems evident that reading efficiency is an important factor in scores on the verbal Scholastic Aptitude Test. Therefore, it is possible to use the verbal Scholastic Aptitude Test scores

as a preliminary screening device for students who need remedial reading in college. It also appears that if the Scholastic Aptitude Test score is included in the final screening, not only would this procedure be justified but the selection of students would be made with greater reliability.

From the results shown above a remedial reading program will apparently have little effect upon courses in English Composition; but in view of the findings of others (7, 9) the results of such a program should be most evident in those subjects usually grouped under the headings of "Social Studies" and "Humanities."

Received July 30, 1945.

References

1. Blommers, P., and Lindquist, E. F. Rate of comprehension of reading: Its measurement and its relation to comprehension. *J. educ. Psychol.*, 1944, **35**, 449-473.
2. *The Cooperative Reading Comprehension Tests*. Cooperative Test Service, 1940.
3. Flanagan, J. C. A new type of reading test for secondary school and college students which provides separate scores for speed of comprehension and level of comprehension. Official Report of the *Amer. Educ. Res. Ass.*, 1938.
4. Flanagan, J. C. A study of the effect on comprehension of varying speeds of reading. Official Report of the *Amer. educ. Res. Ass.*
5. College Entrance Examination Board, 44th Annual Report of the Executive Secretary, 1944.
6. Supplementary information concerning the rating scale.
7. Chauncey, H. The Social Studies Test of the College Entrance Examination Board, 1943.
8. Noyes, E. S. Report on the English Essay Test of the College Entrance Examination Board, 1943.
9. Humber, W. J. The relationship between reading efficiency and academic success in selected university curricula. *J. educ. Psychol.*, 1944, **35**, 17-26.
10. Williamson, E. G. The decreasing accuracy of scholastic predictions. *J. educ. Psychol.*, 1937, **28**, 1-16.

Book Reviews

Steiner, Lee R. *Where do people take their troubles?* Boston: Houghton Mifflin Company, 1945. Pp. xiii + 265. \$3.00.

This is a needed book. Writing in the vein of a reporter, Mrs. Steiner has done "a study of the ways of men and women in trouble." Her subjects are those persons who, whether forlorn, gullible, neurotic, undereducated, unloved, frustrated, or merely idle and bored, are moved to turn for help—or for something—to "the most common public opiates, all of them operating within existing law." The details of the exploitation of these persons and of outright humbuggery which we can infer from the factual records of this book comprise the meat of the book.

Here, preying upon persons in trouble, are a grotesque and sewery array of astrologers, numerologists, "voices," radio counselors, Cosmic guides, Personal columnists, hypnologists, Psycho-Powerhouses, Success specialists, doctors of divine metaphysics—in short the pseudists of every stripe. They ply their lucrative trades without benefit of public license. The author knows them all at first hand and she makes no bones about describing their antics, their audacity, and their piracy. She always writes in good humor, yet with a serious purpose.

Mrs. Steiner was professionally trained as a medical social worker and a psychiatric social worker. Her first move in the study of the psychological underworld was sanctioned by her professional colleagues and by the Illinois Society of Mental Hygiene. In order to secure cases, she listed herself in the "psychology column" of the Chicago telephone directory, classified section, as "The Advisory Service, for professional consultation in the personal, emotional and educational problems of normal people." In response "everyone and anyone came." Next she set up a long-distance mail-order business. Then, after a move to New York, she reversed her strategy. Posing as a prospective client and equipped with various imaginary problems, she visited the whole gamut of New York's super-pseudists. Finally her net to snare people in trouble was put on the air with a weekly radio program.

The result is a clever, well-documented exposé, with human interest sidelights. It is no small accomplishment to describe these professional quacks by name and do it in a way to avoid libel suits. The reader who is temperamentally an optimist may hope that the book will arouse some potential victims of psychological racketeering to realize how foolish and dangerous it is to look for happiness and well-being in such quarters.

Early in her investigation, we infer, Mrs. Steiner got the itch to do "real" counseling herself, and apparently she has kept at it. The book is less explicit on this subject than it ought to be. Despite her crackdown on other "counselors," little is said of her own methods and she does not appear in the slightest degree on the defensive about them. She worked under Alfred Adler and she speaks of her "patients" (p. 135). On p. 137 we read "The Christian Scientists whom I have treated because of their 'religious conflicts' have been suffering from struggles with . . . 'sex.'" On p. 128 it is asserted that "Psychological homosexuality" "can usually be treated successfully."

But the important objective of the book is to point out a vast social problem. Mrs. Steiner urges the need for a nation-wide mental hygiene program which will include radio, extensive use of group (therapeutic) discussions, travelling clinics for small communities, and above all licensed counseling, personal and vocational, on a profession basis. Mrs. Steiner's zeal leads her to proclaim that a "plan for education and professional psychological treatment is our next *must* in governmental interest." (p. 253)

Government, one may suspect, usually helps professions that help themselves, by maintaining high standards of training, service, and publicity.

Richard M. Elliott

University of Minnesota

Gardner, Burleigh B. *Human relations in industry*. Chicago: Richard D. Irwin, Inc., 1945, Pp. xi + 307. \$3.75.

The author, who was for five years in charge of employee relations at Western Electric's Hawthorne Plant, states in his preface that his purpose is "a systematic presentation of the human structure of industry." The book fulfills this purpose. It is a good sociological description of the industrial organization. The structure is logically broken down in terms of line and staff, functional divisions, hierarchical levels, status, sex, age and class differences. These are well described and illustrated with case material.

There is a chapter on personnel counselling in which the case for the non-directive method is summarized but in which there is not enough appreciation of the possible disruptive effects upon the relations between foreman and workers of even the best counselling program. There is an excellently done chapter on merit and incentive wage determination (but somewhat surprisingly no mention of job evaluation), and one on restriction of output. A chapter entitled "Problems of Cooperation" stresses

the potential value of the method of consultative supervision and points to the need for effective communication.

Allowing for the inevitable lack of integration which results from such an approach, the book is well organized. While it does not contain much that has not already been described by the Hawthorne research group, it does bring together in one volume and with some fresh illustrative material much that has been scattered heretofore among a number of books and articles.

The reader with a psychological bias is likely to feel the lack of integration. The industrial organization as a living whole never emerges. In spite of the author's obvious desire to present human relations in a dynamic fashion, one never gets beyond a static impression. The cake has been sliced in a variety of different ways, and one obtains a useful knowledge of what the pieces are like, but the final result is a neat stack of pieces of cake, and no more.

This book points up sharply the lack of a coherent integrated set of principles of organized human effort. We have a wealth of descriptive literature not only of industry, but of church, school and state. Underlying these different forms of organized striving there must be a few—probably rather simple—related principles. What are people trying to do through their organizations? How well do they succeed in doing it? What happens when they fail? What can be done about it? Projected against a framework of generalized answers to these questions, some of the critical problems of our day might become more understandable.

No matter how well one understands the "human structure of industry," the basic problems of human relations in industry remain baffling. A knowledge of structure is helpful, just as a knowledge of anatomy is helpful in medicine. But it is not enough. The author of the book might well argue that we need more and better understanding of the anatomy of industrial organizations. Perhaps we do. But nevertheless the patient is seriously ill. What do we know that will help him to get better?

It requires an almost impossible leap of imagination to go from the kind of description which characterizes this book to the evaluation of policies or an understanding of the important problems of industrial relations. Throughout the book one gets glimmerings of light concerning minute and specific problems. But the big and important problems of organized human effort are left in impenetrable fog.

Perhaps this is asking too much. Perhaps we must proceed slowly and painstakingly to study and understand the industrial organization piecemeal before we will be able to understand and deal with the major issues that confront us today. Certainly this book is a useful adjunct to

the body of descriptive material about industry. Nevertheless, considering the breadth of experience and knowledge of the author, I cannot but profess some measure of disappointment in it.

Douglas McGregor

Massachusetts Institute of Technology

Scheinfeld, Amram. *Women and men*. New York: Harcourt, Brace and Co., 1944. Pp. xv + 453. \$3.50.

A book bringing together from many sciences the relevant material on sex differences has long been needed. In *Women and Men*, a popular book, Scheinfeld has made a more comprehensive survey than is available in any technical book. He goes to the scientific literature in fields such as anatomy, physiology, medicine, genetics, psychology, sociology, anthropology, and vital and economic statistics for his basic material. Each chapter has been read and criticized by scientists qualified in the area covered and contains citations to the appended extensive bibliography. There is little dogmatism in the book. In his preface, Scheinfeld suggests that the reader make a clear distinction between facts and interpretations. The easy style of the book should not blind the reader to the care with which it has been done, even though the scientific reader may wish for more detail. Because powerful attitudes are aroused, whenever sex differences are discussed, some readers will be critical of parts of the book. But this should not blind them to the major job done. Although I could suggest additional material within my own area of interest, I cannot but admire the manner in which its literature has been combed and significant conclusions brought together.

Instead of emphasizing the cross-section material which is, however, adequately covered, Scheinfeld gives a longitudinal developmental picture of the sexes in their social setting and context. It is a long step forward from its predecessor, *Men and Women*, by the late Havelock Ellis (1894). Any comparison of the two will reveal the tremendous advances in the data that have become available, the objectivity with which the problems can now be approached, and the growth of understanding that has come in half a century of scientific effort.

Although the student of applied and industrial psychology will be interested particularly in those chapters which deal with the illnesses, the work, and the social relations of men and women, he will find that the whole book contains implications for the problem of the relation of the sexes in industry and society. Not only should personnel workers, counselors, and guidance experts read it; one can wish that employers and managers, politicians and officials, writers and propagandists, and feminists and traditionists read it. In a sense, this book marks the

end of an era in which some measure of equality of opportunity for the sexes had to be won and the beginning of an era in which both theory and practice are more likely to be based realistically upon scientific data, with full recognition of the principle that men as men and women as women are personality systems affected by and affecting events and relations.

John E. Anderson

University of Minnesota

Radvanyi, Laszlo. *Public opinion measurement*. A survey. Instituto Cientifico De La Opinion Publica Mexicana, 1945. Pp. 88. \$1.00 (U. S. cy.).

Questionnaires were mailed to people with a known interest in public opinion research, most of whom were social scientists and journalists. The twelve questions covered various aspects of the scientific status of public opinion polls, their role in a democratic society, problems of maintaining their integrity, a few technical considerations, and the international cooperation of institutes of public opinion. Although the survey was international in scope, this particular report is confined to the answers received from respondents in the United States.

The first section reports the percentage of respondents giving each answer to each question and in many cases presents the results separately for social scientists and journalists. The second section gives the complete answers of a selected group of respondents for each question.

The report provides very little basis for evaluating the accuracy of the survey. The method of making up the mailing list is not explained. The number of questionnaires mailed is not given, although it is stated that it was a "large number." The problem of bias from a possible selective error in the returns is disposed of with the statement that "the answers received were so numerous and so varied in their origin that they can be considered as quite representative of the opinions of the respective groups."

To the reviewer, many of the questions appear ambiguous and subject to various interpretations. The first question, for example, asks whether public opinion polls can be considered as "scientific method" and "regarded as an important factor in sociological, political, historical, and other research." In the first place, to ask whether any thing in the social sciences can be regarded as "scientific method" invites a variety of answers resulting from differences among people in their definitions of this term. In the second place, it is not clear to the reviewer whether this first question is one question or two: a method might be considered "scientific" and still not be an important factor in the development of the social sciences, and vice versa.

In fairness to this survey, however, it must be admitted that the problems are so involved that the phrasing of questions is necessarily difficult and perhaps the questions are as good as could be expected under the circumstances.

In spite of the limitations of this survey—or at least of the report—the results are of considerable value. The detailed answers of many authorities, given in the second part of the report, are well worth reading although they have been selected and are not necessarily representative. Furthermore, even these selected opinions demonstrate that the problems are important and that there is sufficient diversity of opinion to justify further study.

Probably the spotlighting of these problems was the principal objective of this survey; and, if this is so, it has accomplished its purpose.

Alfred C. Welch

*Knox Reeves Advertising, Inc.,
Minneapolis, Minnesota*

Brandt, H. F. *The psychology of seeing*. New York: The Philosophical Library, 1945. Pp. 240. \$3.75.

Psychology of Seeing is based upon evidence from ocular photography. The author has organized materials derived from his published and unpublished eye-movement researches performed over a period of ten years. Although it is held that eye movements serve as objective symptoms of perceptual processes, it is also contended "that inefficient central processes over a period of time will result in faulty eye-movement patterns which will hinder efficient observation and learning." This emphasis on the importance of "central processes" is to be commended.

After describing the methods employed and certain basic eye-movement tendencies, major sections are devoted to the use of eye-movement photography in evaluating advertising, in the study of learning, and in analyzing responses to art objects. The book is concluded with a discussion of the psychological implications of ocular patterns and a statement of projected studies.

The eye-movement apparatus, designed by the author, is ideally suited for the kind of studies undertaken. With few exceptions the investigations were well planned and the data adequately treated. Interpretations, however, are frequently faulty or inadequate. The author anticipates this criticism by the warning that the reader may at times sense "occasional sweeping statements." Unfortunately there are many typographical errors in both the text and the bibliography.

Miles A. Tinker

University of Minnesota

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to
Donald G. Paterson, Editor, Department of Psychology, University
of Minnesota, Minneapolis 14, Minnesota

Recent occupational trends in American labor—a supplement to occupational trends in the United States. Dewey Anderson and Percy E. Davidson. Stanford University: Stanford University Press, 1945. Pp. 133. \$2.25.

Diagnostic and remedial teaching in secondary schools. Glenn Myers Blair. New York: The Macmillan Co., 1946. Pp. 422. \$3.25.

Moving ahead on your job. Richard P. Calhoon. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 295. \$2.75.

It's how you take it. G. Colket Caner. New York: Coward-McCann, Inc., 1946. Pp. 152. \$2.00.

The application of measurement to health and physical education. H. Harrison Clarke. New York: Prentice-Hall, Inc., 1945. Pp. xvi + 415.

How heredity builds our lives: an introduction to human genetics and eugenics. Robert Cook and Barbara S. Burks. Washington, D. C.: American Genetic Association, 1946. Pp. 64. \$.75.

The executive in action. M. E. Dimock. New York and London: Harper & Brothers, 1945. Pp. ix + 276. \$3.00.

Guidance practices at work. Clifford E. Erickson. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 339. \$3.25.

Occupations: a selected list of pamphlets. Gertrude Forrester. New York: The H. W. Wilson Co., 1946. Pp. 240. \$2.25.

Trends in employment and earnings for nineteen graduating classes of a teachers college. John S. French. New York: Teachers College, Columbia University, 1945. Pp. vi + 103. \$1.85.

Vocational aptitude tests for the blind. Samuel P. Hayes. Watertown: Perkins Publications, Perkins Institution and Massachusetts School for the Blind, 1946. Pp. 32. \$.25.

Human leadership in industry: challenge of tomorrow. Sam A. Lewisohn. New York and London: Harper & Brothers, 1945. Pp. ix + 112. \$2.00.

Psychology in industry. Norman R. F. Maier. Boston: Houghton Mifflin Co., 1946. Pp. 463. \$3.00.

- Group psychotherapy: A symposium.* J. L. Moreno. New York: Beacon House, 1945. Pp. 305. \$5.00.
- Signs, language and behavior.* Charles Morris. New York: Prentice-Hall, Inc., 1946. Pp. 365. \$5.00.
- Psychology.* Normal L. Munn. Boston: Houghton Mifflin Co., 1946. Pp. 497. \$3.25.
- The adolescent in social groups.* Frances Burks Newman. Stanford University: Stanford University Press, 1946. Pp. 94. \$1.25 p. \$2.00 cl.
- Youth, marriage and parenthood.* Lemo D. Rockwood and Mary Ford. New York: John Wiley and Sons, Inc., 1945. Pp. 279. \$3.00.
- Evaluation of group guidance work in secondary schools.* Georgia May Sachs. Los Angeles: The University of Southern California Press, 1946. Pp. 120. \$2.50.
- Marriage and the family.* Edgar Schmiedeler. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 285. \$1.80.
- Collective bargaining.* Leonard J. Smith. New York: Prentice-Hall, Inc., 1946. Pp. 416. \$3.75.
- The dynamics of human adjustment.* Percival M. Symonds. New York: D. Appleton-Century Company, 1946. Pp. 674. \$5.00.
- The psychology of normal people.* Revised. Joseph Tiffin, Frederic B. Knight, and Easton Jackson Asher. Chicago: D. C. Heath and Co., 1946. Pp. 550. \$3.00.
- Hitler's professors. The part of scholarship in Germany's crimes against the Jewish people.* Max Weinreich. New York: Yiddish Scientific Institute—Yivo, 1946. Pp. 291. \$3.00.
- Controlled eye movements versus practice exercises in reading.* Frederick Lowell Westover. New York: Bureau of Publications, Teachers College, Columbia University, 1946. Pp. 99. \$1.95.
- Proceedings of the third annual visual education institute.* W. A. Wittich. Madison: University of Wisconsin Summer Session, 1945. Pp. 114.

Journal of Applied Psychology

EDITED BY: DONALD G. PATERSON, UNIVERSITY OF MINNESOTA

Consulting Editors

PAUL S. ACHILLES, *Psychological Corporation*; WALTER V. BINGHAM, *A.G.O., War Department*; HAROLD E. BURTT, *Ohio State University*; ARTHUR I. GATES, *T. C. Columbia University*; JOHN G. JENKINS, *University of Maryland*; IRVING LORGE, *T. C. Columbia University*; QUINN MCNEMAR, *Stanford University*; WILLARD C. OLSON, *University of Michigan*; JAMES P. PORTER, *Swarthmore, Pennsylvania*; EDWARD K. STRONG, JR., *Stanford University*; MORRIS S. VITELES, *University of Pennsylvania*; JOSEPH ZUBIN, *N. Y. Psychiatric Institute*.

Table of Contents

<i>Studies in Supervisory Evaluation: Q. W. FILE AND H. H. REMMERS</i>	421
<i>Studies in Job Evaluation 5. An Analysis of the Factor Comparison System as it Functions in a Paper Mill: C. H. LAWSHE AND R. F. WILSON</i>	426
<i>The Learning Curve for Flying an Airplane: W. N. KELLOGG</i>	435
<i>The Purdue Mechanical Adaptability Test:</i>	
C. H. LAWSHE, I. A. SEMANEK, AND J. TIFFIN	442
<i>The Relative Readability of Newsprint and Book Print:</i>	
D. G. PATTERSON AND M. A. TINKER	454
<i>Age of Starting to Contribute versus Total Creative Output: H. C. LEHMAN</i>	460
<i>The Use of the Harrower-Erickson Multiple Choice Rorschach Test With a Selected Group of Women in Military Service: M. C. WINFIELD</i>	481
<i>The Relationship Between Subjective Estimates of Personal Adjustment and Ratings on the Bell Adjustment Inventory: J. TUCKMAN</i>	488
<i>Item Difficulty of Some Wechsler-Bellevue Subtests:</i>	
A. I. RABIN, J. C. DAVIS, AND M. H. SANDERSON	493
<i>The Relationship Between Knowledge of Human Development and Ability to Use Such Knowledge: J. E. HORROCKS</i>	501
<i>Keysort Method of Scoring the Minnesota Multiphasic Personality Inventory:</i>	
M. P. MANSON AND H. M. GRAYSON	509
<i>Profile Analysis of the Minnesota Multiphasic Personality Inventory in Differential Diagnosis: PAUL E. MEEHL</i>	517
<i>The K Factor as a Suppressor Variance in the Minnesota Multiphasic Personality Inventory: PAUL E. MEEHL AND STARKE R. HATHAWAY</i>	525
<i>Book Reviews</i>	565
<i>New Books, Monographs, and Pamphlets</i>	573

Published Bi-monthly by The American Psychological Association, Inc.

Prince and Lemon Sts., Lancaster, Pa., and
1515 Massachusetts Ave., NW, Washington 5, D. C.

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the Act of March 3, 1879

Copyright, 1946, by The American Psychological Association, Inc.

Journal of Applied Psychology

Vol. 30, No. 5

October, 1946

Studies in Supervisory Evaluation

Quentin W. File and H. H. Remmers

Purdue University

As management starts examining the general quality of the supervision which was obtained during the war, it becomes increasingly apparent that improved selection technics would pay their costs many times over. Good "bosses" do not just happen. Unfortunately too, poor ones cannot be completely rehabilitated by intensive company training programs. Interviews, though an absolute necessity, cannot, when used alone, provide complete objective evaluations of potential supervisors' qualifications.

The purpose of this article is twofold: First, to compare the findings of Sartain's study,¹ as reported in the August issue of this Journal, with the findings reported in the study by File² which resulted in the development of the test, *How supervise?*, and second, to report more recent evidence of the validity of this test.

Briefly summarizing some of the points reported by Sartain, it would seem that in his study:

1. Rather reliable ratings of the abilities of the supervisors studied were obtained.
2. All the standardized tests considered measured factors other than those included in the ratings obtained on these supervisors.
3. Mental ability as measured by Tiffin and Lawshe's *Adaptability test* is in that particular plant negatively related to the attitudes and understandings set forth by the test, *How supervise?*, as being necessary for supervisory success.

Before raising questions as to the general applicability of these findings, it should be pointed out that Dr. Sartain was very careful to em-

¹ Sartain, A. Q. Relation between scores on certain standard tests and supervisory success in an aircraft factory. *J. appl. Psychol.*, 1946, 30, No. 4 (1946).

² File, Quentin W. The measurement of supervisory quality in industry. *J. appl. Psychol.*, 1945, 29, 323-337.

phasize that his data were obtained in a single war plant. He likewise mentions the possibility that his criterion may possess inherent weaknesses. The following observations, therefore, should be considered, not as a criticism of his study, but as an attempt to evaluate the possibilities for generalization which obtain.

1. In the last stages of World War II, the aviation industry was probably much less stable than the average established manufacturing organization. Its payrolls contained a sizable proportion of workers drawn from other more permanent industries. Many of these workers by this time realized that loss of their jobs in the near future was a foregone conclusion. Its top management, too, may well have reflected the effects of rapid expansion by failing to formulate adequate standards for supervisory performance.

2. Management's ratings may be reliable without being valid. Management above the operating level must rely primarily on line organization channels for its information. Ratings by higher management may well be a reflection of the immediate supervisor's opinions, thus producing spuriously high correlations.

3. The reported correlations between management's ratings and the ratings of job success were higher than the reliabilities of the ratings themselves (as estimated by the Spearman-Brown Prophecy Formula). Since this would indicate some systematic bias in the scores, it seems possible that all the criterion scores may have a common origin, namely the judgment of the individual giving the grades for job success.

Sartain and File agree that multiple ratings by line management do not correlate to any significant extent with scores on the experimental edition of *How supervise?* File in his study of some 577 supervisors in ten industries also failed to find relationships above .15 between Management's ratings and: (1), work experience; (2), education or supervisory training; (3), supervisory experience, and (4), stability of employment.

One highly significant difference between Sartain's study and previously reported evidence is the relationship between general mental ability and scores on the experimental edition of *How supervise?* Sartain reports of correlation of $-.44$ ($N = 40$) where general mental ability was measured by the *Adaptability test*. File, using a slightly different approach, obtained a correlation of $.35$ ($N = 577$) between highest educational level attained and scores on the supervisory ability test. Since there is a known positive relationship between general mental ability and scholastic achievement, the findings appear contradictory. The following observations are submitted in support of File's findings: (1), Tests requiring reading ability normally correlate .30 or more with "intelligence" test scores; and (2), File's study was based on over five hundred

supervisors in ten different industrial concerns while Sartain's study included only forty supervisors in an expanded war industry. This difference between the findings of the two studies may constitute reason for questioning whether the supervisors in the sample studies by Sartain are sufficiently typical to be used as cases for drawing general conclusions concerning the usefulness of tests in selecting supervisory personnel.

Recent Evidence of the Validity of *How supervise?*

Considerable evidence has now been obtained that the *revised edition* of *How supervise?*³ does possess validity for the selection of industrial supervisors and a more comprehensive validation program is now in progress. Results reported by three industries may be summarized as follows:

I. One form of the test was given to 46 successful supervisors and 14 non-supervisors (by-passed because of judged lack of ability) in an office machine manufacturing company. Results of the investigation are as follows:

	Successful Supervisors	Non-Supervisors
Per cent Above 50th Percentile	80	15
Per cent Below 50th Percentile	20	85
Average Percentile Score	75	23
Critical Ratio of Difference Between Proportions Above and Below 50th Percentile Point	5.8	

II. Excerpts from report by the Supervisor of Testing in a company which manufactures surgical supplies.

"1. I have computed the reliability of the test on 50 cases and find that its reliability is .85. A reliability of .85 means roughly that in 60 cases out of 100 a person will achieve approximately the same score when tested a second time.

"The conventional odds-evens method of determining the reliability was used.

"2. I have done some work on validity using an expedient method which is not fully acceptable, but it shows a positive trend which encourages me to continue this work when more valuable criteria are available. As a result of my evaluations, I selected either present supervisors or those whom I had recommended as potential supervisors, and considered these men successful supervisors. Since my evaluations had been

³ Published by The Psychological Corporation, 522 Fifth Avenue, New York City.

made independent of the *How supervise* test score, I felt that using this as a standard I could select another group whom I had not recommended for comparison. Using the same method, I selected 20 people who, in my opinion, substantiated by test results, could not become supervisors at . . . (Name of Company). I contrasted the two groups. The following table will give you my findings:

	<i>N</i> = 20 Superior Group	<i>N</i> = 20 Inferior Group
Mean	54	38
Standard Deviation	9.	11.
Standard Error of Mean	2.01	2.45
Critical Ratio	4.4	

"The above brief table shows you that the average for the 'super-visors' is higher than the 'non-supervisors.' The standard deviation shows that the 'supervisor' group is more consistent than the 'non-supervisor' group. The Standard Error of the mean shows that in 68 cases out of 100, the average of the superior group will range between 56.01 and 51.99 and that in 68 cases out of 100, the average for the inferior group will range between 40.55 and 35.45. The Standard Error of the difference between the two means is 4.4 which is sometimes called the *critical ratio*. If the critical ratio is 3, it is considered significant and means that in 100 chances out of 100 the test is differentiating. 4.4 is additional assurance that it is really differentiating."

III. Report submitted by the General Manager of a relatively large laundry. Sixteen supervisors were given *How supervise*. These individuals were divided into the following groups:

- Group I. The "Company has complete confidence in Group I to handle all types of supervisory problems. Each individual is rated by us as superior in this respect.
- Group II. "This group is doing an excellent job, but occasionally needs follow-up on practices. Delicate situations occasionally require assistance.
- Group III. "These individuals are new, and have only recently been assigned supervisory responsibilities. Preliminary evidence would indicate success.
- Group IV. "These usually give substandard supervisory performance.
- Group V. "Experience has proved that these people can be given only the most limited supervisory responsibilities. However, both have other qualities so valuable to us that their services will be retained."

	Number in Each Group	Average Raw Score	Percentile Value of Average Score
Group I	6	54.7	96
Group II	3	49.7	91
Group III	3	44.7	79
Group IV	2	40	68
Group V	2	32	33

Obviously these studies do not constitute conclusive evidence of the universal validity of *How supervise?* as a supervisory selection device. The numbers of cases in the studies were small. Selection of the criterion groups was made either by a personnel department man or by some unexplained method. These criterion groups by definition constitute the extremes of the distribution and could be expected to show more startling differences than if all the supervisory personnel in those concerns were included.

On the other hand, evidences of satisfactory discrimination were found in three widely different supervisory groups. Though the numbers of individuals considered in each sample were small, statistically significant differences were obtained. Considering all three studies as a unit, 116 supervisors or potential supervisors were measured and, if statistically-combined, the computed discrimination ratio would be higher than those reported for the individual studies.

Summarizing briefly, the independent indications of the validity of *How supervise?* which have been obtained to date are:

1. Significant increases in supervisory understanding have been measured by administering the test before and after supervisory training.
2. As reported above, significant differences have been found between "successful" supervisors and individuals by-passed because of lack of supervisory ability.
3. No company has reported evidence, either subjective or statistical, that differences in supervisory ability are not measured by the revised edition of the test. A considerable number of concerns have accepted it validity on the basis of subjective evidence and report receiving valuable results.

With this background of strong preliminary evidence of validity, a more comprehensive investigation is now being undertaken. Provision is also being made for the construction of a new business and management form of the test. This edition is intended for use among management above the foreman level and among supervisors in business organizations as well as those in industrial concerns.

Received June 24, 1946.

Studies in Job Evaluation. 5. An Analysis of the Factor Comparison System as it Functions in a Paper Mill *

C. H. Lawshe, Jr., and R. F. Wilson

Division of Applied Psychology, Purdue University

With the increasing acceptance of job evaluation as a wage negotiation and stabilization technique come several fundamental problems. What basic judgment factors determine the differential wages to be paid workers on various jobs? Can these factors be discovered and systematized into a list which will cover all or nearly all jobs, so that differential wage levels can be determined fairly and objectively? The purpose of this series of papers is to work toward an answer to these questions by analyzing some of the currently accepted job evaluation systems and their results as they function in various industrial situations.

Factor analysis techniques in previous studies have yielded two factors which define themselves with striking similarity and consistency from plant to plant and sometimes a third which is related to the uniqueness of the type of plant or class of jobs being rated. The first factor correlates highly with mental and skill requirements and other job elements which connote these qualities and has been designated as "Skill Demands." The second factor correlates highly with such elements as working conditions, and hazards and has been designated as "Job Characteristics."

It has been found in the former studies that an abbreviated scale composed of selected items from the original scale will yield results correlating highly with the original scale ratings. The resultant displacement of jobs in terms of cents per hour, which would occur as a result of using the abbreviated scale, would be so small as to suggest strongly the advisability of simplifying job evaluation systems to attain administrative efficiency and lower cost in the operation of the job evaluation program.

* This article is a "prior publication," the author paying complete costs. The scheduled 80 pages per issue is thereby increased by the corresponding amount, thus the "early publication" of this article is a direct contribution to the subscribers of the *Journal of Applied Psychology* without handicap to those authors whose articles are accepted and printed in their regular turn.

Purpose of This Study

The former studies have all dealt with point rating systems. It was felt that a study of the Factor Comparison System, as presented by Bengé, Burk, and Hay, (1), might yield particularly significant results for two reasons.

First, the Factor Comparison System, through its method of job to job comparison, should tend to minimize the "halo" effect. Under the point rating systems, the analyst, in considering a job description, then rating it on a series of scales, might tend to rate that job at the same level on several scales. For instance, the rater, having decided that a high degree of mental ability was required for a job, might tend to let that decision influence his rating on several subsequent scales, such as Responsibility, Versatility, Experience, etc. This would increase the "halo" effect, and would tend to give spuriously high correlations of the job elements with each other, which in turn would result in a smaller number of factors and distorted weighting of the job elements on those factors.

Second, the Factor Comparison System might avoid to some degree the tendency of a rater to check many jobs at or near the same level on a given scale. If a rater tends to rate all jobs at or near the middle of an item scale, say Working Conditions, the effect is to eliminate any discrimination on the basis of that job element, producing the same effect as if that element had been left out of the scale altogether.

On the basis of these considerations, this particular study promised to yield further insight into the nature of the basic job evaluation factors as identified in the judgment process.

The Factor Comparison System

Selection of Key Jobs. The Factor Comparison System of job evaluation involves the comparison of jobs being rated with a scale of "key" jobs rather than the evaluation of jobs against an *a priori* point scale. Fifteen to twenty five jobs, ranging from the highest paid jobs to the lowest paid jobs in the plant, are selected by a job evaluation committee. These must be jobs which have clearly defined job descriptions and the rates of which are generally judged to be fair.

Ranking of Key Jobs. Members of the committee, individually and collectively, rank these key jobs in order of difficulty on Mental Requirements. Then they rank the key jobs again on Physical Requirements, and so on until the key jobs have been ranked on each of the following five job elements: Mental Requirements, Physical Requirements, Skill Requirements, Working Conditions, and Responsibility.

The Salary or Wage Breakdown. The wage or salary for each of the key jobs is then analyzed by estimating the amount of it that is being

paid for each of the job elements. The following example will serve to illustrate. Shop Clerk,—Job Description No. 124: Mental Requirements, \$23; Physical Requirements, \$9; Skill Requirements, \$20; Working Conditions, \$16; add Responsibility, \$32; making a total salary of \$100.

Matching the Breakdown With the Rank. The amounts estimated by the committee as being paid for Mental Requirements on each of the key jobs are entered on a sheet opposite the final ranking, as illustrated by the following hypothetical example:

Rank	Job Title	Estimated Amount Paid for Mental Requirements
1	Dept. Supervisor	\$70
2	Asst. Supervisor	47
3	Machine Bookkeeper	31
4	Card Punch Operator	21
5	Mail Clerk	25
6	File Clerk	18
7	Messenger Boy	11

When the matching is done, some of the amounts (as No. 5 above) may be out of line with the ranking. If the amount is only slightly out of line, the committee may decide to reprice the elements on that job to adjust this difficulty; otherwise that job is eliminated from the list of key jobs. This matching of amounts and rankings is then completed for each of the five scales in this manner, and only those jobs which fall in line are retained as key jobs.

Rating the Bulk of the Jobs. At this point the job analyst has five "measuring sticks" with which to rate other jobs. Each level on the "measuring sticks" is described by a representative job, and each level also has its "price tag." In evaluating a different job, the analyst first compares that job with the key jobs on the Mental Requirements scale to determine which of the key jobs demands a similar degree of mental ability. Having selected a point on the scale at or near the most comparable key job, the analyst assigns the indicated amount to be paid for Mental Requirements on the job being rated. By repeating this process on each of the five scales, the analyst has five amounts which added together equal the indicated salary or wage for the job being rated.

Procedure

The cooperating industry, a paper mill, furnished Factor Comparison job evaluation data on one hundred and seventy-six job classifications.

The cents-per-hour values obtained for the five job elements on each job classification were punched on IBM (machine sort) cards. The constants determining the intercorrelations were obtained from machine tabulations, the correlations computed, and a correlation matrix prepared (Table 1).

Table 1

Intercorrelation Coefficients Between Ratings on Five Factors and Total Points for 176 Jobs in a Paper Mill

	2. Mental Require- ments	3. Physical Require- ments	4. Skill Require- ments	5. Respon- sibility	6. Total
1. Working Conditions	-.09	.76	-.03	-.10	.23
2. Mental Requirements		-.16	.92	.91	.92
3. Physical Requirements			-.06	-.16	.16
4. Skill Requirements				.89	.94
5. Responsibility					.90

Factor analysis of these intercorrelations by Thurstone's centroid method yielded two factors with job element loadings as shown in Table 2. The extraction process was discontinued when Thurstone's *phi* test was satisfied. These two factors account for virtually all of the variability in total points since the communality (h^2) equals unity. Rotation was accomplished by the graphical method described by Guilford (2). The Wherry-Doolittle shrinkage selection method was applied to the values in the correlation matrix, and three items were selected for an abbreviated scale.

Identification of Factors

Factor I. Inspection of Table 2 demonstrates a convincing affirmation of the two basic factors found in the former studies. Factor I

Table 2

Factor Loadings for Five Scale Items and Total Points

Scale Items	Before Rotation			After Rotation		
	k_1	k_2	h^2	k_1	k_2	h^2
Working Conditions	.37	-.77	.73	.06	+.85	.73
Mental Requirements	.82	.54	.96	.96	-.20	.96
Physical Requirements	.31	-.78	.70	.00	+.84	.70
Skill Requirements	.87	.46	.97	.98	-.11	.97
Responsibility	.81	.53	.94	.95	-.20	.94
Total Points	.98	.22	1.00	.99	+.16	1.00

clearly has to do with Mental Requirements, Skill Requirements, and Responsibility, as demonstrated by the heavy loadings .96, .93 and .95. Working Conditions and Physical Requirements have practically no weighting on Factor I. The correlation matrix (Table 1) shows the high correlation of these three job elements with each other. Evidently the raters are judging only slightly different aspects of the same thing. These three job elements correspond convincingly with the definition of the "Skill Demands" factor developed in the previous studies.

Factor II. Factor II quite clearly defines itself in terms of the other two job elements, Working Conditions and Physical Requirements, and

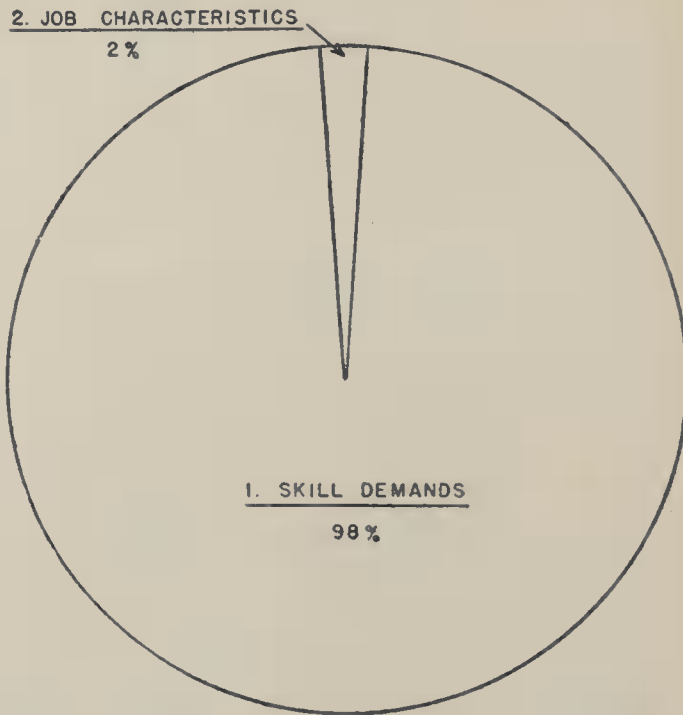


FIG. 1. Graph showing proportion of the total variability attributable to each of the two factors.

this corresponds with the "Job Characteristics" factor obtained in the previous studies. The small negative loadings which the other three elements have with Factor II are reasonable to expect. The correlation matrix (Table 1) further shows that Working Conditions and Physical Requirements correlate much higher with each other than with any of the other job elements.

Proportion of Variability Accounted for by Each Factor. As in former studies, Factor I accounts for an extremely high proportion of the final variation of job rates. Since the partial correlation (k_1) of Total Points with Factor I is .99, the coefficient of determination would be k_1^2 which

equals .98, and Factor I would account for 98% of the variability in Total Points as shown in Figure 1. In like manner it is seen that Factor II accounts for 2% of the variation in Total Points.

Adequacy of an Abbreviated Scale

Items Selected. Application of the Wherry-Doolittle shrinkage selection process to the values in Table 1 using total points as the criterion selected three of the job elements for an abbreviated scale. Since Skill Requirements alone correlates .94 with Total Points, the procedure was followed to identify those items which would increase the correlation most. The process was discontinued after selection of the third element because the multiple correlation of Skill Requirements, Working Conditions, and Mental Requirements with the results of the original scale had already become slightly higher than .99. Addition of a fourth element would have increased the multiple correlation by less than .008. The multiple correlation results are shown in Table 3.

Table 3

Multiple Correlation Coefficient Between Groups of Items and Total Points

Items	<i>R</i>
Skill Requirement Alone	.94
Skill plus Working Cond.	.97
Skill plus Working Cond. plus Mental Req.	.99

Accuracy of the Abbreviated Scale. The wage administrator, however, is less interested in abstract correlation figures, and more interested in the number of cents per hour by which the various jobs would be displaced as a result of using the abbreviated scale. Figure 2 shows the "scatter" obtained by plotting the wages, as calculated from the abbreviated scale, against the job wages obtained originally with the complete scale. Still further analysis of the data, as shown in Table 4, reveals that 166 out of the 176 jobs would be displaced by four cents or less, and that no job would be displaced by more than 8 cents. The average difference in wage rate is 1.7 cents.

Discussion

In regard to the advisability of using an abbreviated number of elements in the job evaluation system as opposed to use of the system in its original and complete form, there is another important consideration beyond merely demonstrating that an abbreviated scale will yield results

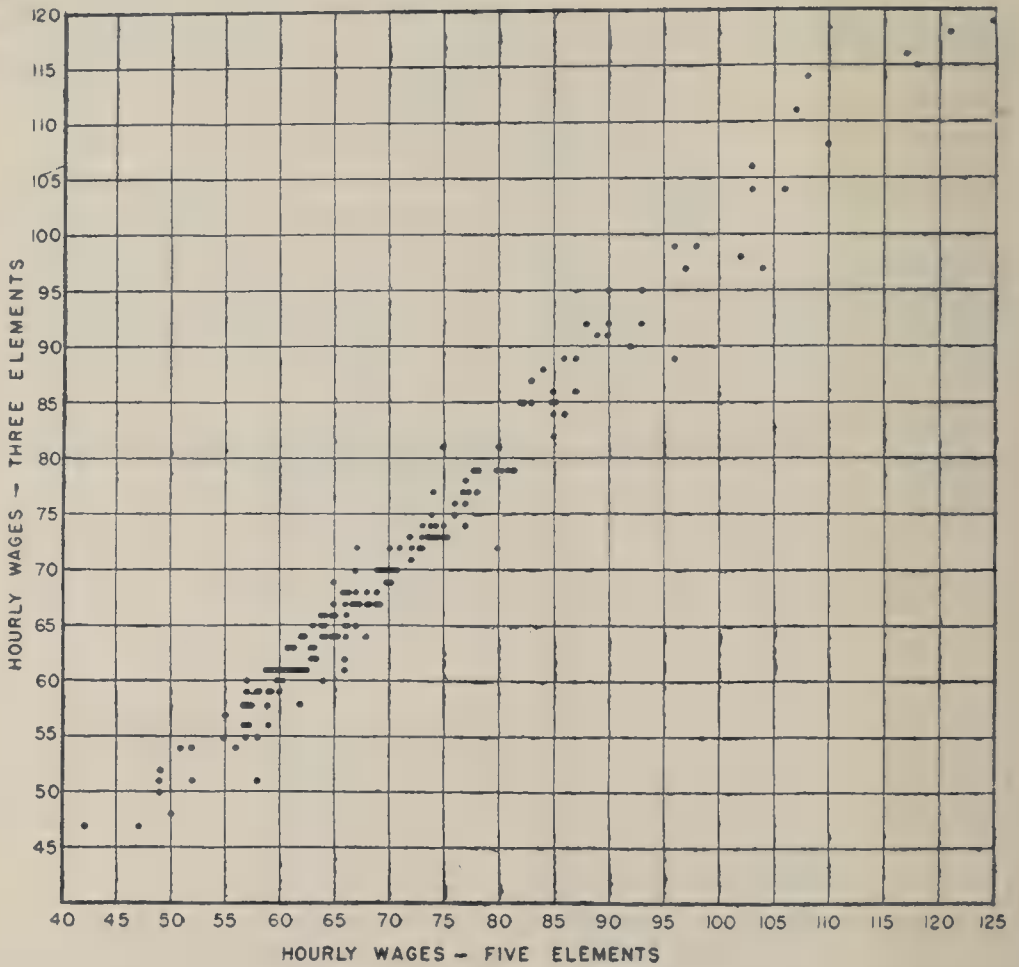


FIG. 2. Graph showing wages computed by the abbreviated three element scale plotted against wages computed by using the original five-element scale.

Table 4

Differentials in Cents per Hour Between Rates Computed with Five Factors and Rates Compiled from Three Factors

Difference in Cents	Frequency	Frequency Percentage	Cumulative Frequency	Percentile Values
0	32	18	32	18
1	66	38	98	56
2	41	23	139	80
3	16	9	155	88
4	11	6	166	94
5	4	2	170	96
6	2	1	172	98
7	3	2	175	99
8	1	1	176	100

closely comparable to the original system. No job evaluation system yet devised seems to have attained perfection, and thus these abbreviated systems are being compared with a fallible criterion. Bearing in mind that the criterion is fallible, and that the simplified or abbreviated systems do yield closely comparable results, the question arises: are the ratings obtained through the use of the abbreviated system perhaps as good as, or better than, the ratings obtained from the use of the original system? For instance, if all of these job evaluation files were to be withheld from the members of this job evaluation committee, and they were to go through the complete job rating procedure again, would it not be possible that rates on the various jobs would vary from the original values more than the abbreviated system results vary from the original? Chances are great that they would vary more, for the results of the abbreviated scale correlated .99 with the rates obtained from the original scale. Probability is that the reliability of the original system is not that high. The pertinent question then becomes: Are not the abbreviated scale ratings as good as or better than the original scale ratings?

Since there is no theoretically perfect criterion of job evaluation, it seems that the question, "Which ratings are the best?", would have to remain unanswered. However, it is well to bear in mind that job evaluation as an industrial wage administration technique does not eliminate the chance and error inherent in human judgment, but merely attempts to set a framework in which these human judgments may work more systematically and reliably. Reliability, then, may be an extremely important aspect of job evaluation and should be subjected to systematic investigation.

Summary and Conclusions

Factor Comparison System job evaluation data from a paper mill were subjected to the Thurstone Factor Analysis technique following the intercorrelation of points awarded on each of the job elements and the total. Rotation was accomplished by the graphical method. The Wherry-Doolittle shrinkage selection method was used to select three of the job elements for an abbreviated scale. The following findings are supported:

1. The Factor Comparison Job evaluation system, which through its mechanics should tend to force the rater to make five separate and distinct ratings of the jobs, actually effected judgments on two principal axes in this industrial situation.

2. The first principal axis, or factor, has a heavy loading in mental and skill requirements (and other job elements, such as responsibility,

which connote these qualities). This factor called "Skill Demands" is responsible for 98% of the final variation in job rates.

3. The second principal axis, or factor, has heavy loadings in Physical Requirements and Working Conditions. This factor called "Job Characteristics" accounts for only 2% of the final variation in job rates.

4. Application of the Wherry-Doolittle shrinkage selection method selected three of the five job elements for an abbreviated scale. These three elements, Skill Requirements, Working Conditions and Mental Requirements, when combined correlate .99 with the original scale.

5. While the reliability is not known, it is probable that the correlation between the original and the abbreviated scale is as high as could be obtained with the existing, reliability and that the abbreviated scale can be considered as valid and as usable as the original scale.

Received June 10, 1946.

References

1. Bengt, E. J., Burk, S. L. H., and Hay, E. N. *Manual of job evaluation*. New York: Harper and Brothers, 1941.
2. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill Book Co., Inc., 1936, pp. 502-505.
3. Lawshe, C. H., Jr., and Satter, G. A. Studies in job evaluation. 1. Factor analysis of point ratings for hourly paid jobs in three industrial plants. *J. appl. Psychol.*, 1944, 28, 189-198.
4. Lawshe, C. H., Jr. Studies in job evaluation. 2. The adequacy of abbreviated point ratings for hourly-paid jobs in three industrial plants. *J. appl. Psychol.*, 1945, 29, 177-184.
5. Lawshe, C. H., Jr. Studies in job evaluation. 3. An analysis of point ratings for salary paid jobs in an industrial plant. *J. appl. Psychol.*, 1946, 30, 117-128.
6. Lawshe, C. H., Jr., and Alessi, S. L. Studies in job evaluation. 4. Analysis of another point rating scale for hourly-paid jobs and the adequacy of an abbreviated scale. *J. appl. Psychol.*, 1946, 30, 310-319.

The Learning Curve for Flying an Airplane *

W. N. Kellogg

Indiana University

The object of the present investigation was to examine the process of learning to fly. It was directed specifically towards the plotting of learning curves and the study of the manner in which the student-pilot eliminates his incorrect or erroneous responses as he masters the flying technique.

Procedure

Apparatus. In order to obtain objective records, a special mechanism, known as the pilot-response recorder, was developed and installed in a Piper Cub Trainer. This device is illustrated in Figure 1. The pilot-response recorder weighs about 10 pounds and makes automatic graphic tracings of the extent and duration of the rudder, aileron, and elevator movements while the plane is in flight. By means of a system of cams or wedges (*W*, Fig. 1) the absolute extent of the airplane control movements is transmitted to the clockwork polygraph of the pilot-response recorder in direct linear proportion. The writing pointers are mounted on sleeves and move in a straight line across the paper. Errors which are common in similar devices, such as the distortion introduced by the arcs of writing levers which are pivoted at a fulcrum, errors of changing air pressure within pneumatic systems, or the variation in the elasticity of tambours at different tensions, were eliminated by this method. The entire apparatus was mounted in a concealed position behind the cockpit.¹ It was therefore possible to keep the student-pilot from knowing that records of his flying were being made at all.

Sample records made by this device are reproduced in Figure 2. The lines show movements of the rudder, elevator, and ailerons which were traced during the process of making landings. The first ground contact in each instance is indicated by the vertical broken line, so that, except for subsequent bumps, the portion of each tracing to the right of the

* The investigation was financed by the Civil Aeronautics Authority through the Committee on Selection and Training of Civilian Pilots of the National Research Council. The data for this study were obtained in 1939 and 1940, but publication was necessarily withheld until after the termination of the war.

¹ The pilot-response recorder has been patented by Indiana University under the name of the airplane multiple control recorder.

broken line represents taxiing. Time intervals shown on the bottom horizontal line are 10 seconds in length.

Types of Records Made. A standard course, which required about 10 minutes to fly, was laid out with fixed pilons on the ground. The course included four left turns and three right turns. Pilot-response records for flying the course with records of the corresponding landings and take-offs were made by *both student and instructor* after approximately every 30

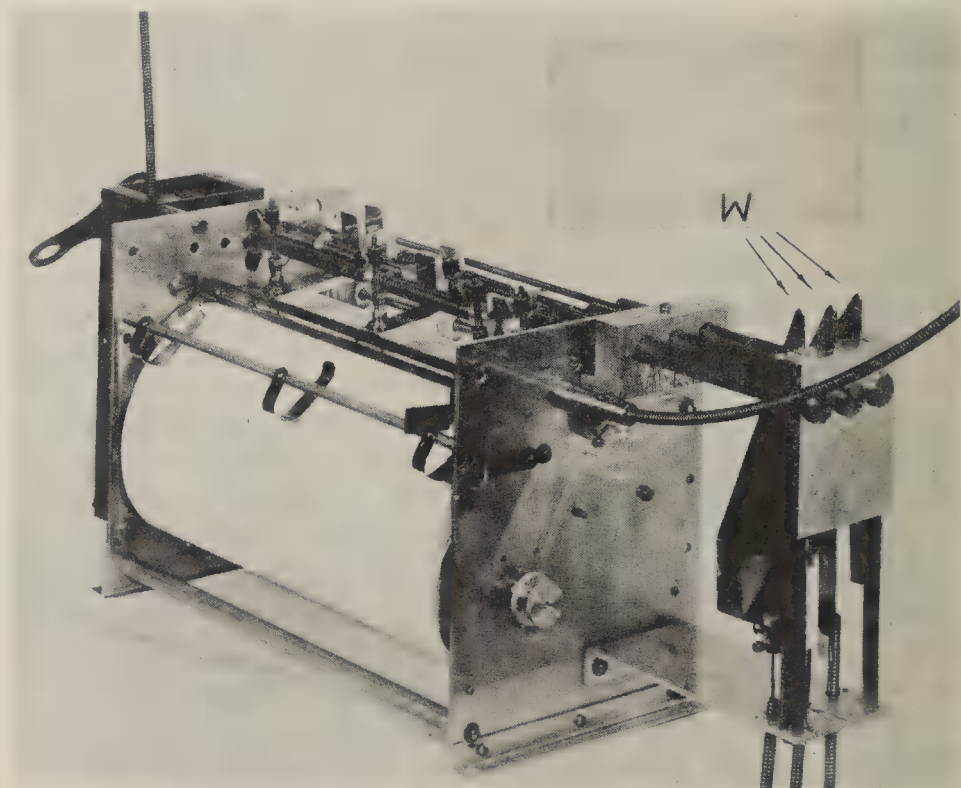


FIG. 1. The pilot-response recorder is a light-weight polygraph by means of which the movements of the airplane controls can be graphically traced. A patented system of cams or wedges (*W*) transmits the control movements to the paper in linear proportion to their absolute extent. Errors which might be introduced by the arcs of writing levers which are pivoted at a fulcrum, by pneumatic systems, or by the variable tensions of tambour diaphragms are eliminated by this construction.

minutes of flight instruction. Periodic records were also taken of steep and shallow figure-eights and of 360 degree glides-to-a-landing.

The Weather-control Technique. The object of having the instructor make flying records along with the student was to obtain some kind of a base or standard with which to compare the student's performance. This base could not be a fixed one, but would be constantly changed or modified by variable weather conditions. To cancel out this possible source

of error, the instructor made the same maneuvers as the student, either immediately before or immediately after the student had made them. Since the student's and the instructor's records were obtained but a few moments apart, over the same terrain, the difference between them could be regarded as a difference between the skill of the expert or finished pilot and the performance of the beginner.

Every student record, therefore, had paired with it the corresponding record made by the instructor under the same flying conditions. To find

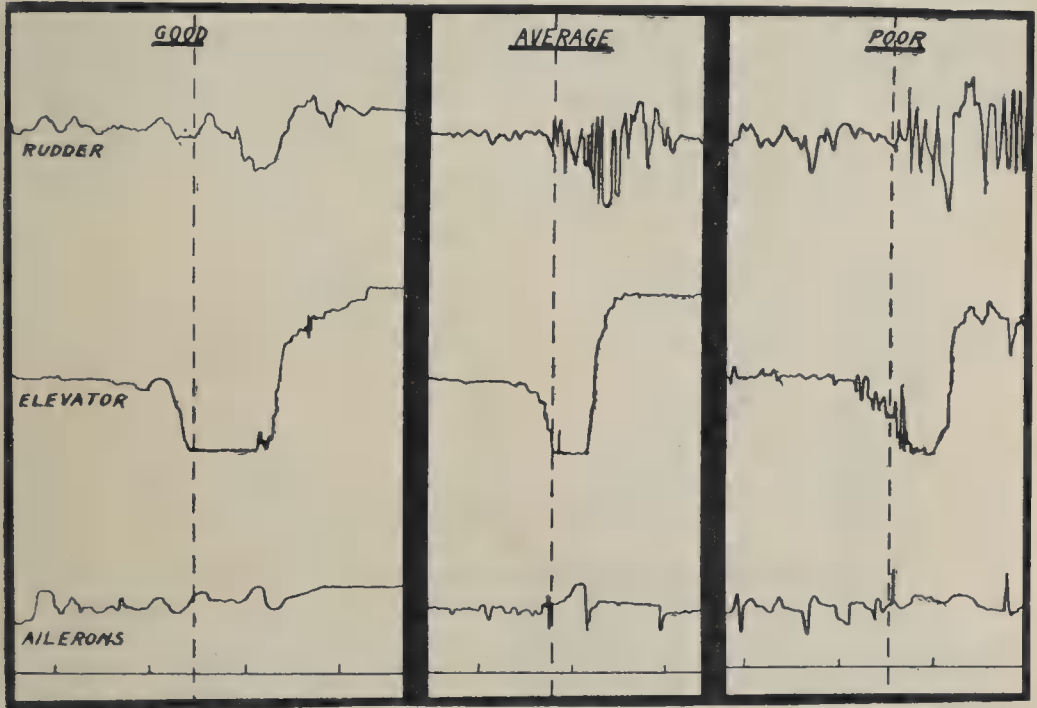


FIG. 2. The irregular lines show movements of the rudder, elevator, and ailerons which were recorded during the process of making landings. First ground contact in each instance is indicated by the vertical broken line. The tracings to the right of the broken lines, therefore, represent taxiing. Time intervals on the horizontal line at the bottom are 10 sec. in length.

what a student's errors were one compared the objective record of his flight with the appropriate control record made by the instructor. This method has been called the weather-control technique.

Quantifying the Data. The graphic records made by the pilot-response recorder were measured by means of a special device known as a graphometer, which automatically totals the vertical deflections or oscillations from the horizontal of any irregular or wavy line.² Readings from the

² W. N. Kellogg. A device for measuring kymographic records. *J. exp. Psychol.*, 1936, 19, 383-385.

graphometer converted to numerical form the total amount of movement of each of the airplane controls within any given time period. By comparing the graphometer readings of the student and the instructor it was possible to tell at once which person moved any given control more or less than the other person, and *exactly how much more or less* he moved it.

Results

The results in this report cover the training of two student-pilots. Presented below are a few selected items which seem to offer the most promise for the analysis of the learning process.

Course Records. In Figure 3 is shown the learning curve plotted from graphometer readings of the elevator movements of student-pilot C,

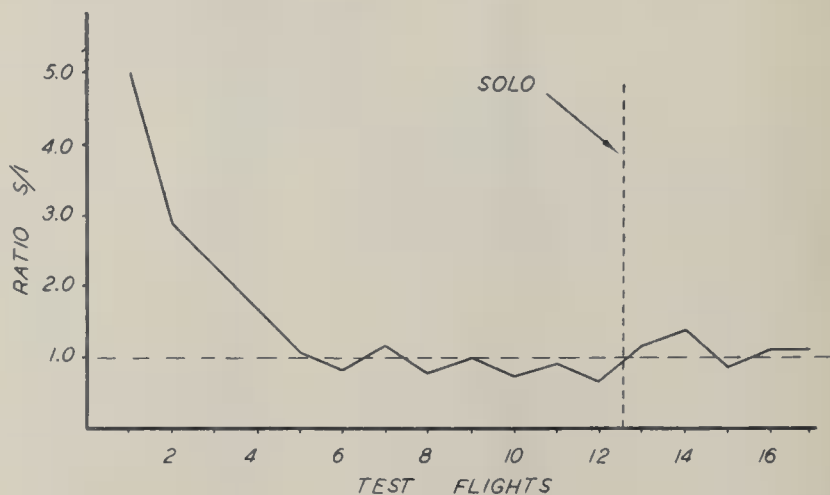


FIG. 3. Learning curve plotted from elevator movements of subject C, showing the gradual elimination of overcontrolling with practice in flying over a standard course. Correct manipulation of the controls is represented by the horizontal line.

during his flights over the standard ground course. The points plotted are all ratios of the amount of elevator movement made by the student (S) divided by the amount of elevator movement made by the instructor (I). The curve includes 17 test flights or, roughly, 12 hours of instruction (17 half-hour periods plus 17 ten-minute periods of course flying). Student C made his first solo flight between test flight numbers 12 and 13.

Since the points on the graph are all ratios, one can tell at once that student-pilot C began his flying by moving the elevator about five times as much as the instructor moved it. He was therefore overcontrolling very badly. A ratio of 1.0 (indicated by the broken horizontal line) would mean that the student moved the elevator the same amount that the instructor did within the same time period. It will be seen from Figure

3 that student C gradually eliminated his elevator overcorrections so that, after the eighth test flight, he was not far from the instructor's performance.

In Figure 4 is shown a similar curve for student-pilot C, but one which is a composite or combination of the movements of all three of the airplane controls. It appears from this learning curve that the student-pilot on the whole moved the controls less than the instructor moved them. This is indicated by the fact that the level of the curve is most of the time below the ratio of 1.0. Comparing the first part of the learning curve in Figure 4 with the first part of the curve in Figure 3 one may infer that since subject C overcontrolled so much with the elevator he must have undercorrected with the other controls. As a matter of fact, this individual was much too limited in his rudder movements, as the graph of the course records for the rudder (not presented here) demonstrated.

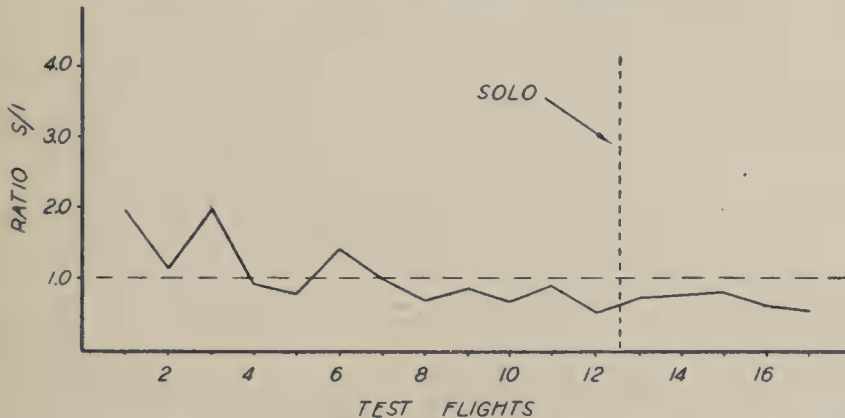


FIG. 4. Learning curve showing reduction in the manipulation of all three controls as compared to correct or ideal use of controls which is indicated by horizontal line.

Records of Landings. One of the most difficult maneuvers which the new pilot has to perfect is the maneuver of landing. It is, moreover, a maneuver from which many records can be easily obtained and one which must remain highly practiced with the pilot as long as he flies an airplane. It should be clear also that in the maneuver of landing the elevator plays by far the most important part. A good landing is actually made only with the elevator and throttle (unless flaps are used). The rudder and ailerons should not be employed except in the approach to the field and in correcting for bumps in rough air.

The perfect landing is one in which the stick is gradually drawn backwards (the tail lowered) as the plane loses speed in its landing glide. In the case of a three-point landing the stick should be all the way back at the moment the tail and landing wheels make contact with the ground

(see Fig. 2). Poor landings are those in which there is too much *forward* movement of the stick. The student "pumps" the stick back and forth as he tries to "find the ground." Improvement in landings should therefore be shown by the reduction in forward stick-movements with practice.

In order to get at this problem, pilot-response elevator records were measured for a period of 15 seconds as the plane came into a landing. "A landing" was arbitrarily defined by this means as the 15 seconds of flying time which ended with ground contact. The learning curve plotted from such measurements, combined from the graphometer readings of the elevator movements of two subjects (C and P), is shown in Figure 5. Each point on the graph is a ratio of forward movements (F) divided by

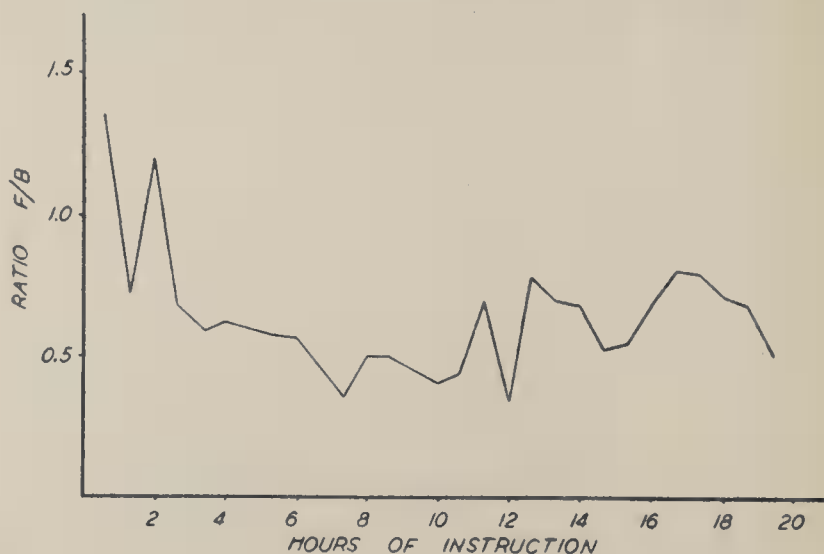


FIG. 5. Showing improvement in the use of the elevator during landings only. Composite learning curve for two subjects.

backward movements (B) of the stick—combined for two student pilots. When the ratio is high (1.0 or 1.5) it means that the subject is pushing forward too much on the stick during his landings. When the ratio is low it means that he is making few forward movements and that the landings are therefore "good."

From an examination of Figure 5 it appears that there is a rapid improvement in landing skill for the first few hours of instruction, and that thereafter the progress is slow—as in the mastery of any difficult skill.

Conclusions

The following propositions seem to be justified by the limited data of this study.

1. The objective analysis of airplane control movements can be used to show progress in the development of flying skill.

2. By means of the pilot-response recorder and the graphometer, the psychologist can tell which controls the pilot is manipulating incorrectly, in which direction his errors occur, and how great they are.

3. The weather-control technique seems to be adequate to cancel out variations in flying conditions.

4. Learning curves for various maneuvers in flying are essentially the same as those obtained in the development of other skills.

5. There is no evidence of plateaus in the curves of learning to fly, as plotted from the present data.

Received October 18, 1945.

The Purdue Mechanical Adaptability Test *

C. H. Lawshe, Jr., Irene A. Semanek, and Joseph Tiffin

Division of Applied Psychology, Purdue University

Personnel administrators are realizing more and more that modern personnel programs are strengthened and improved through the use of personnel tests. Although many tests are available for use in industry, there is still a definite need for tests designed specifically for that purpose. *The Purdue Mechanical Adaptability Test*¹ is a test consisting of 60 questions about practical mechanical facts which are answered "yes," "no," or "don't know." It was designed to measure "knack" in mechanical, electrical, and related activities by means of an evaluation of experience in these areas.

In the process of standardizing and validating the test on industrial populations the personnel departments of eleven manufacturing concerns cooperated by administering this test to applicants for jobs of a mechanical nature and to employees already working on such jobs. In the case of employees presently on the job, success ratings were obtained from supervisors.

The intent of the authors was to select items of high internal consistency and at the same time to select items relatively unrelated to mental ability. Both of these objectives are important since highly consistent items may also have a high relationship to intelligence.

Construction and Standardization

Original Construction. The initial step in the development of *The Purdue Mechanical Adaptability Test* consisted of the construction of 400 items of practical information in seven work areas. A deliberate effort was made to have these items meet the following criteria:

1. The item must deal with practical information obtainable from first hand contact and not with theoretical principles or concepts.

* This article is a "prior publication," the author paying complete costs. The scheduled 80 pages per issue is thereby increased by the corresponding amount, thus the "early publication" of this article is a direct contribution to the subscribers of the *Journal of Applied Psychology* without handicap to those authors whose articles are accepted and printed in their regular turn.

¹ *The Purdue Mechanical Adaptability Test* by C. H. Lawshe, Jr., and Joseph Tiffin is copyrighted by the Purdue Research Foundation and is distributed by the Division of Applied Psychology, Purdue University, Lafayette, Indiana.

2. The item must be as short as possible.
3. All words (except technical vocabulary) should fall within the ability of a normal eighth grade student.

These 400 items were arranged into four test forms, ST, UV, WX, and YZ, which were administered in mimeographed form to 138, 110, 109 and 122 male students respectively in grades 10, 11, and 12 in two comprehensive high schools. *The Adaptability Test*,² a test of general mental ability, was administered to these populations at the same time and the 30% scoring highest and the 30% scoring the lowest were segregated. The "Kelley technique" of item validation described by Lawshe³ was employed and *D*-value (discrimination values) on *The Mechanical Adaptability Test* items were computed using these mental ability groups as criterion groups. Since it was desirable to discard items that were known to be highly related to intelligence, arbitrary *D*-value limits of $+.3$ and $-.3$ were established and all items having *D*-value outside of that range were discarded. This left 207 items which were distributed among the four preliminary forms as follows: ST, 53; UV, 51; WX, 45; and YZ, 58.

Item Selection. Only these selected items were scored, and the high 30% and low 30% of the population on each form were identified to be used as criterion groups for computing *D*-values against these total scores. By this process, only those items having a *D*-value of $.8$ or better were retained. One hundred items were then selected from these four preliminary forms, each item of which had a *D*-value of $.8$ against total scores on the selected items and none of which had a *D*-value computed against mental ability that deviated from zero by more than plus or minus $.3$. These 100 items constituted Form R and were arranged in approximate order of difficulty. This trial form, also mimeographed, was administered to 250 boys in the 10th, 11th, and 12th grades in a trade school and to 189 male college students, a substantial number of whom were engineering students. Members of this latter group were asked to make written criticisms of any of the items that in their judgment were in any way ambiguous. On the basis of these comments, minor adjustments in items were made and the revised test was printed as Form S. Further analysis involved items that were common to Form R and to Form S of the test.

Revision. Form R and Form S of the *Purdue Mechanical Adaptability Test* were then administered to men in industrial situations. Of the 462

² Tiffin, Joseph, and Lawshe, C. H., Jr. The Adaptability Test: a fifteen minute mental alertness test for use in personnel allocation. *J. appl. Psychol.*, 1943, 27, 152-163.

³ Lawshe, C. H., Jr. A nomograph for estimating the validity of test items. *J. appl. Psychol.*, 1942, 26, 848-849.

cases used in the final selection of test items, 364 were industrial applicants in a steel mill, and 98 were already on the job in the following types of plants: foundry, screw manufacturing, electrical supply manufacturing, and farm machinery manufacturing. *The Adaptability Test* designed to measure mental alertness was also administered to all but twenty-four of the cases.

The "Kelley technique" of item validation already mentioned was again used for estimating the validity of individual test items. Since it was desirable to reduce the number of items by selecting those that tended to measure the same thing and hence were highly consistent and yet at the same time not highly related to mental ability, internal and external criteria for item selection were again used.

The internal criterion was the total score on the 100 items of the test prior to the item analysis; the external criterion was the score obtained on *The Adaptability Test*. In computing internal consistency *D*-values, criterion groups consisted of the 30% making the highest scores on all hundred items and the 30% making the lowest scores on the same items. After *D*-values were computed, papers for those cases for which *The Adaptability Test* scores were available were isolated and criterion groups selected to include the 30% scoring highest and the 30% scoring lowest.

A scattergram of internal consistency *D*-values vs. mental ability *D*-values was then plotted. All items which did not yield an internal consistency *D*-value of .5 or better were discarded. Since it was desirable to minimize the intelligence factor in the test, all items yielding a *D*-value of .7 or greater against *The Adaptability Test* were discarded, thus eliminating those items which tended to be associated most closely with mental ability.

This process yielded 60 items which were incorporated into Form A (Men) with the items arranged in order of increasing difficulty. Table 1 shows the content areas of the questions incorporated in Form A and the number of items in each area.

The method of scoring the earlier forms as well as Form A is as follows: (1) the number of correct responses is counted; (2) that total is doubled; (3) to that is added the number of responses marked "Don't Know." This yields the raw score⁴ which may be converted to a per-

⁴ This raw score is based upon a modification of the standard correction for guessing ($R - W$) where two choices exist. The derivation is as follows:

$$\text{Score} = R - W + 60$$

$$W = 60 - R - DK$$

$$\text{Score} = R - (60 - R - DK) + 60 \text{ or } 2R + DK$$

The modification has two advantages: (1) since 60 is added to each score, all negative scores are eliminated, and (2) in combination with a scoring stencil, scoring is simpler since no marks need be made on the test papers and wrongs need not be counted.

Table 1

Content Areas Incorporated in the Purdue Mechanical Adaptability
Test Form A (Men)

Area	No. of Items
Woodwork and Finishing	10
Automobile	17
Electricity and Radio	18
Machine Shop	4
Plumbing	4
Sheet Metal	2
Miscellaneous	5
Total	60

Table 2

■ The Relationship of Test Scores on Form A of the Purdue Mechanical Adaptability
Test and Other Measures

Test or Measure	<i>N</i>	<i>r</i>	σ_r
California Capacity: Non-language	25	.41	.13
California Capacity: Language	25	.12	.16
Bennett Test of Mechanical Comprehension	33	.71	.09
Minnesota Paper Formboard	39	.18	.16
Age	40	.32	.14

Table 3

Differences between Means for College Sub-groups and their Significances

Group	<i>N</i>	Mean	C.R.
Mechanical and Aeronautical Engineers	71	103.1 \pm 1.4	6.3
Science, Pharmacy, and Physical Educ.	103	91.7 \pm 1.2	
Mechanical and Aeronautical Engineers	71	103.1 \pm 1.4	3.8
Civil, Metallurgical, and Electrical Engineers	54	95.6 \pm 1.5	
Civil, Metallurgical, and Electrical Engineers	54	95.6 \pm 1.5	2.1
Science, Pharmacy, and Physical Educ.	103	91.7 \pm 1.2	
All Purdue Students	274	99.7 \pm 0.7	

Table 4
Percentile Norms for the Mechanical Adaptability Test, Form A

Percentile	College Men*				
	Industrial Men (N = 1015)	"Non-Engineering" (N = 103)	"Non-Mechanical" (N = 54)	"Mechanical" (N = 71)	All (N = 274)
100	116	114	114	118	118
95	108	109	111	115	114
90	105	107	111	112	111
80	99	102	108	110	108
70	95	97	105	108	106
60	90	93	101	106	103
50	86	91	97	104	99
40	81	89	94	102	96
30	76	86	91	99	93
20	72	82	87	94	89
10	65	76	80	88	82
5	61	72	76	83	77
1	57	69	72	79	74

* College groups were constituted as follows: "Non-engineering" included Science, Pharmacy, and Physical Education students; "non-mechanical" included Civil Engineers, Electrical Engineers and Metallurgical Engineers; "Mechanical" included Mechanical Engineers and Aeronautical Engineers; the "all" classification was a random sample of Purdue University students.

centile equivalent by means of Table 4. The highest possible score on Form A is 120, which is obtained when all 60 items are correct.

Validity

The validity of the *Purdue Mechanical Adaptability Test* for industrial use in employee and trainee allocation is shown by the relationship between scores on the test and employee success on the job as measured by supervisory ratings. The studies which follow involved employees who took either Form R or Form S of the test. However, in the results presented here, only those items which now appear in Form A were scored. The first three of the validity studies below involved cases not included in the primary group used in the item selection procedure.

Ice Company Mechanics. The correlation⁵ between scores on *The Purdue Mechanical Adaptability Test* and success on the job as measured by supervisor's rankings of fourteen experienced mechanics in an ice company was $.86 \pm .07$. The effectiveness of this test in identifying the highest rated mechanics can be seen by examining the scattergram in

⁵ All plus and minus error designations in this paper pertain to standard errors.

Figure 1. The best mechanic was the highest scorer on *The Mechanical Adaptability Test*, and the poorest one scored lowest. The other mechanics tend to fall in a linear pattern.

Further inspection of Figure 1 shows that if a score of 90 were the minimum acceptable score for this job, 80% of the less desirable mechanics (rated 1 and 2) would be rejected and only 20% would be accepted. Of the more desirable mechanics (rated 3 or higher), 89% would be accepted and only 11% would be eliminated. If the minimum acceptable score were set at 95, all of the least desirable mechanics would fall below this score, as would 60% of the average (rated 3); but at the same time, all of the more desirable mechanics (rated 4 or 5) would fall at or above this critical score. This indicates that a mechanic scoring higher on *The Mechanical Adaptability Test*, other things being equal,

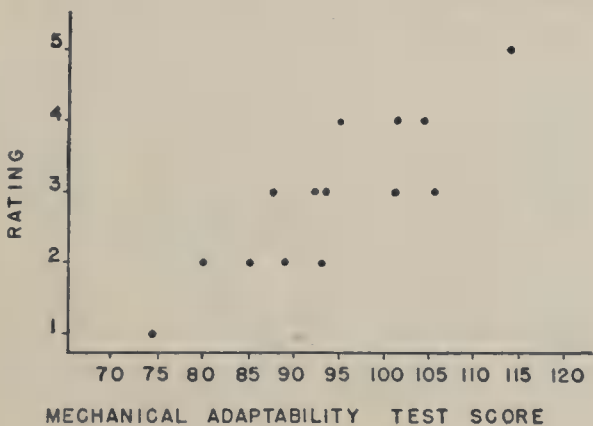


FIG. 1. A scattergram of scores on Form A and job success ratings of fourteen mechanics in an ice company.

would on the average be more successful on this job than one scoring lower on the test.

Time Study Men. Six time study men from a company manufacturing musical instruments were ranked by the supervisor and *The Purdue Mechanical Adaptability Test* was administered to them. The rank order correlation was found to be $.75 \pm .18$. A scattergram of these cases is shown in Figure 2. It will be noted that there is one inversion in the case of the time study man rated 4 and that otherwise there is a perfect correlation. Although there were only six cases in the study, it is interesting to note that the three highest scorers were also the three highest ranking employees in the group.

Steel Mill Apprentices. Twelve apprentices in a steel mill took the test at the time of hiring. Figure 3 shows a plot of test scores and rank order ratings made by the supervisor after they had been on the job. This group included the following apprentices; four machine shop, two

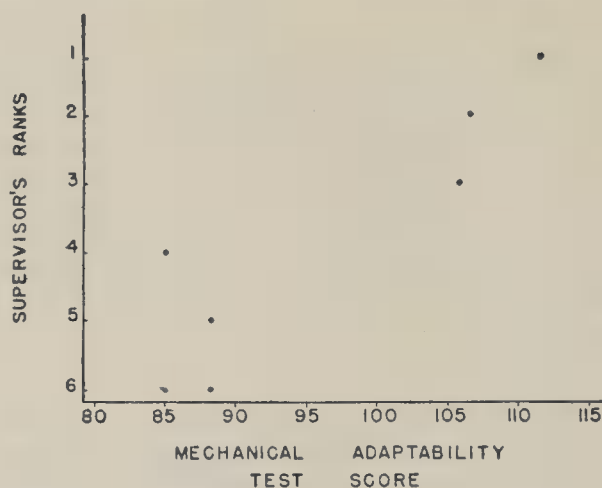


FIG. 2. A scattergram of test scores and supervisory rankings of six time study men in a musical instrument plant.

masonry, and one each of electrical, maintenance, carpenter shop, blacksmith, pipe shop helper, and electrical construction apprentices. The rank order correlation was found to be $.39 \pm .24$. When the apprentices rated one to six are, arbitrarily, put into the "high" group and the others in "low," 50% are in the "high" group to begin with. With a minimum acceptable score of 70, the percentage of people rated high is increased to 55%; a minimum score of 89 increases the percentage of "high" scores to 60%. While this relationship between scores is not statistically significant, in combination with the other studies reported here, it yields some indication of what can be expected.

Foundry Workers. Data on another group of twelve men who were employed in a foundry were studied. Ratings were prepared by the

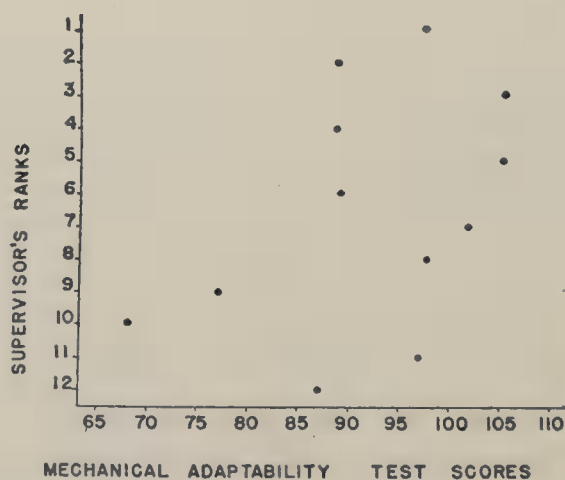


FIG. 3. Test scores and rankings of twelve steel mill apprentices.

foreman after the employees had been working in the plant thirty days. Three men were rated very good, eight as fair and good, and one as poor. An inspection of the data showed that if a critical score of 70 were used, the employee rated poor and two of the fair to good class would not be acceptable, yet all those rated very good scored above this critical point.

Screw Manufacturing Plant. The test was administered to 46 experienced machine operators in a screw manufacturing plant. The operators were rated 5.0 (lowest rating), 5.5, 6.0, and up to 19.5 (highest). Ratings and scores were plotted on a scattergram. Figure 4 shows that if mechanics rated 13.5 or better are considered "high," the percentage of employees rated "high" increases as the minimum acceptable test score is increased. The figure shows that with a minimum acceptable score of 75, about 44% of those passing were "high" rated. With a minimum

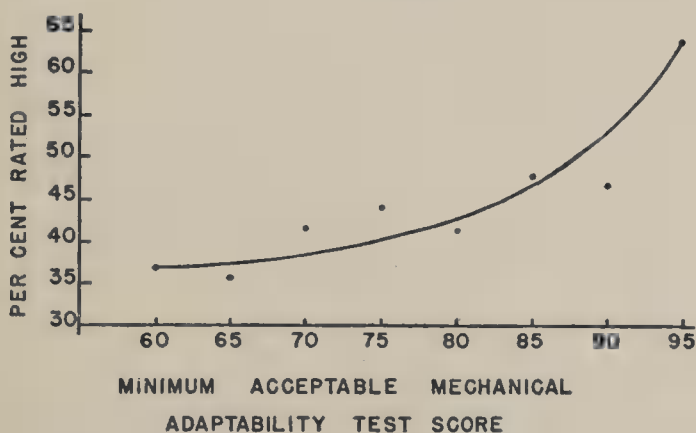


FIG. 4. Graph showing percentage of screw machine operators who are rated "high" when successively higher critical test scores are employed.

acceptable score of 85, 48% of those passing are "high," and with a critical score of 95, 64% are in the "high" rated group. Since only 37% of the whole group were rated "high" to begin with, the number of "high" rated people increases as the minimum acceptable score is raised.

Electrical Company Apprentices. A group of 40 trade apprentices from an electrical equipment manufacturing plant consisted of fourteen machinists, eleven toolmakers, seven diemakers, one foundryman, and seven miscellaneous electrical workers. The apprentices were rated C+, B-, B, B+, A-, and A (highest). Since there was a tendency for apprentices in some of the fields to be consistently rated higher than apprentices in other fields, the apprentices in each group were divided into approximately the highest rated half and the lowest rated half. Those rated high in each of the trade areas were pooled in the computations. Figure 5 shows that 47% of the apprentices in all trade areas were rated

"high." However, with a critical score of 80, 50% are rated "high"; and when higher critical scores are considered, the percentage of "high" rated employees tends to increase.

Validity for Guidance Use. The adequacy of differential group norms as an estimate of test validity has frequently been questioned. However, so marked were some of the student group differences obtained that the facts are recorded here for such value as they may have. Men students, 274 in number, in the Elementary Psychology course at Purdue University were given Form A. These men, mostly underclassmen, were drawn from all curricula in the University. By combining students in certain related programs or curricula, three sub-groups each exceeding 50 in number were obtained. For example, those students enrolled in Science,

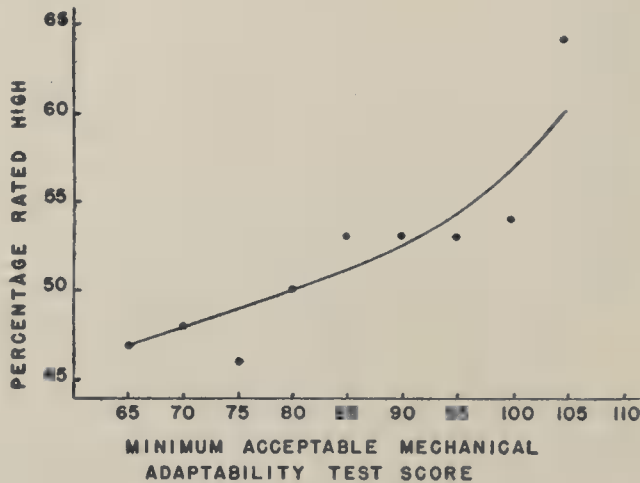


FIG. 5. Graph showing percentage of trade apprentices in an electrical company who are rated "high" when successively higher critical scores are employed.

Pharmacy, and Physical Education were combined. Those students in specific engineering curricula were divided into the "mechanical" group (Mechanical Engineering, Aeronautical Engineering, and Air Transportation) and the "non-mechanical" group (Civil Engineering, Metallurgical Engineering and Electrical Engineering). Table 3 presents the mean scores for each group as well as the critical ratios based upon the differences between them. All obtained differences are in the expected direction, the "mechanical" group being at the top and the non-engineering group at the bottom. Note that the critical ratios range from 2.1 to 6.3. These facts coupled with the industrial validity data just presented might be useful in evaluating the test for guidance purposes.

Test Characteristics

Reliability. Reliability of Form A was obtained by the split-half method with two population samples, an industrial group and a college group. The first sample (not the primary group used for item analysis) consisted of 487 men, all applicants for industrial jobs with an optical manufacturing plant. Scores for the odd numbered items yielded a coefficient of $.84 \pm .01$ when correlated with scores on the even numbered items (stepped up by Spearman-Brown formula). The same procedure was repeated with a new group of 201 college men and a corresponding coefficient of $.80 \pm .03$ was obtained.

Relationship to Mental Ability. Since *The Purdue Mechanical Adaptability Test* was designed to predict success on mechanical jobs, one of the aims was to eliminate items closely identified with intelligence. One study of the intelligence factor was made with the 485 cases on which the split-half reliability was computed. Total scores on Form A were correlated with scores on *The Adaptability Test*. A coefficient of correlation of $.32 \pm .04$ was obtained. It should be kept in mind that this was not the primary group and that the range of mental ability was quite great. Using a group of 173 college men, the correlation between *The Adaptability Test* and *The Purdue Mechanical Adaptability Test*, a coefficient of $.17 \pm .07$ was found.

The raw scores on the *Otis Self-Administering Higher Examination* of twenty-five mechanics employed by a farm machinery manufacturing firm correlated $.08 \pm .20$ with scores on Form A.

Language and non-language raw scores on *The California Capacity Test* were available on forty apprentices varying in training experience from six months to three and one-half years. The group was in training with an electrical manufacturing plant, and included machinists, tool-makers, die makers, and electrical testers. Table 2 shows the relationship of the language and non-language factors to *The Purdue Mechanical Adaptability Test*. An examination of the standard errors of r for the two parts of the test as shown in Table 2 indicates a considerably greater probability that there is a true correlation in the case of the non-verbal test. This is evidence that a certain communality exists between the *Purdue Mechanical Adaptability Test* and the non-verbal test of mental ability that does not exist in the case of the verbal mental ability test.

Relationship with Other Measures. Other data available on the machinist apprentices mentioned above included raw scores on the *Bennett Test of Mechanical Comprehension*, *The Minnesota Paper Form-board*, and the ages of the apprentices. Correlations with scores on Form A of *The Mechanical Adaptability Test* are indicated in Table 2. Of

special interest is the correlation of .71 between Form A of *The Purdue Mechanical Adaptability Test* and the *Bennett Test of Mechanical Comprehension*, since it purports to measure a similar trait or ability.

Summary

Forms R and S of *The Purdue Mechanical Adaptability Test* were administered to 1,015 industrial applicants or employees already on the job in eleven manufacturing plants. The "Kelley technique" as described by Lawshe was used for estimating the validity of test items. A revision based on internal and external criteria resulted in Form A of the test for men or boys. The split-half method of determining the reliability was utilized and stepped up by the Spearman-Brown formula. Validity studies were made by using as criteria the job success of employees as estimated by supervisor's ratings. The following results were obtained:

1. Of the 100 items included in Forms R and S, the 60 items yielding an internal consistency *D*-value of .5 or better, and an external consistency *D*-value of .6 or less against *The Adaptability Test* were selected for Form A. Thus, items which tended to be associated with mental ability were eliminated; those that were highly consistent were retained.

2. Validity studies of test scores and job success as estimated by supervisor's ratings were made on employees in six different manufacturing plants. A study of the data showed that in general when the minimum acceptable score on *The Mechanical Adaptability Test* was increased, an increase in the per cent of people rated "high" could be expected, thus increasing the proportion of desirable employees on the job.

3. An examination of mean scores of special college curricular groups revealed significant differences from group to group in the expected direction.

4. The reliability of the revised test as obtained on a secondary group of 487 cases by the split-half method and stepped up by the Spearman-Brown formula was $.84 \pm .01$. The corresponding value for a group of college men was found to be $.80 \pm .03$.

5. The item analysis procedures succeeded to a large degree in accomplishing the objective of formulating a test that is relatively unrelated to intelligence as measured by various standard intelligence tests. This is indicated by the following coefficients of correlation with various intelligence tests: *The Adaptability Test*, $.17 \pm .07$ and $.32 \pm .04$; *Otis Self-Administering Higher Examination*, $.08 \pm .20$; *The California Capacity Test: Language* $.12 \pm .16$; *The California Capacity Test: Non-language*, $.41 \pm .13$.

6. In one sample, correlations of $.71 \pm .09$ and $.18 \pm .16$ with the *Bennett Test of Mechanical Comprehension* and the *Minnesota Paper Form-board* were obtained.

7. *The Purdue Mechanical Adaptability Test* is useful in identifying men or boys who are mechanically inclined and are likely to succeed on jobs of a mechanical nature and it can be used in personnel situations as a supplement to regular employment procedures.

Received June 19, 1946.

The Relative Readability of Newsprint and Book Print *

Donald G. Paterson and Miles A. Tinker

University of Minnesota

In earlier studies^{1, 2, 3} the authors have investigated the readability of newsprint and of book print but no direct comparison has been made between the two kinds of printing. There are, however, various hints that newsprint may be read at a slower rate than book print. Paterson and Tinker³ found a consistent tendency for 6 and 8 point book type to be read slower than larger sizes of type. The most frequently used type size in newspaper printing is 7 and 8.² In another kind of study, Tinker⁴ discovered that in reading 7 point newsprint a greater intensity of light was needed for adequate perception than was necessary with 10 point book type.⁵ Nevertheless, since newsprint and book print represent somewhat different typographical situations, there is not enough evidence for an adequate statement of their relative readability.

A direct comparison of the two kinds of printing is made in this study. Specifically, the purpose of the investigation is to compare the speed of reading commonly used newsprint and book print.

In our survey of newspaper printing,⁶ the following was the most common practice for body types: Ionic type face was most frequently used, with Opticon the most popular of the newer type faces; 12 pica line width; 7 and 8 point type; and one point leading. In the same study we noted that one point leading improves readability of newsprint but that two point gives no added advantage. In view of these results and practices we chose the following newspaper typography for use in this study: Arrangement number one was 7 point Ionic No. 5 in a 12 pica line

* Grateful acknowledgment is given to the Graduate School, University of Minnesota, for research grant to finance this study.

¹ Tinker, M. A., and Paterson, D. G. Differences among newspaper body types in readability. *Jour. Quart.*, 1943, 20, 152-155.

² Tinker, M. A., and Paterson, D. G. War time changes in newspaper printing practice. *Jour. Quart.*, 1944, 21, 7-11.

³ Paterson, D. G., and Tinker, M. A. *How to make type readable*. New York: Harper and Brothers, 1940, pp. 209. (Obtainable from the authors.)

⁴ Tinker, M. A. Illumination intensities for reading newspaper type. *J. educ. Psychol.*, 1943, 34, 247-250.

⁵ Tinker, M. A. The effect of illumination intensities upon speed of perception and upon fatigue in reading. *J. educ. Psychol.*, 1939, 30, 561-571.

⁶ Tinker, M. A., and Paterson, D. G., *op. cit.*, 1943.

width with one point leading. Arrangement number two consisted of 8 point Opticon in a 12 pica line width with one point leading. Both were printed on newsprint paper stock. Incidentally, Opticon was the most readable type face of nine investigated in another study.⁷ For the

Cheltenham Book Type:

10 point with two point leading

26. James' fountain pen went dry when he was doing his homework for school. He was very cross because until he got some more glue he could not continue his work. 27. The boys saw coming towards them an old woman, bent with sorrow, dressed in deepest black. They thought, turning from their play to watch her pass, how happy she looked. 28. On

Opticon Newsprint:

8 point with one point leading

26. James' fountain pen went dry when he was doing his homework for school. He was very cross because until he got some more glue he could not continue his work. 27. The boys saw coming towards them an old woman, bent with sorrow, dressed in deepest black. They thought, turning from their play to watch her pass, how happy she looked. 28. On

Ionic No. 5 Newsprint:

7 point with one point leading

26. James' fountain pen went dry when he was doing his homework for school. He was very cross because until he got some more glue he could not continue his work. 27. The boys saw coming towards them an old woman, bent with sorrow, dressed in deepest black. They thought, turning from their play to watch her pass, how happy she looked. 28. On Sunday Mr.

FIG. 1. Samples of book type and newsprint type used in study of relative readability.

book print we chose an optimum typographical arrangement. (See Paterson and Tinker, *op. cit.*, 1943.) This consisted of Cheltenham type face, 10 point with two point leading in a 20 pica line width on eggshell paper stock. Samples of the printing used are shown in Figure 1.

⁷ Tinker, M. A., and Paterson, D. G., *op. cit.*, 1943.

The reading material consisted of Forms A and B of the Chapman-Cook Speed of Reading Test. Although performance on Form B is equivalent to that on Form A on the average, this is not always true for small samples.⁸ A control group was introduced, therefore, to check on this equivalence. There were 30 paragraphs of 30 words each in each test form. The reading time allowed for each form was $1\frac{3}{4}$ minutes.

Three groups of 90 college students each served as subjects. In Group I (control) the subjects read book print in Form A and Form B. In Group II, Form A was book print and Form B was the 8 point Opticon newsprint. And in Group III, Form A was book print, and Form B was the 7 point Ionic No. 5 newsprint. In addition to the above comparisons, an additional 117 college students ranked samples of the print according to apparent legibility and according to pleasingness. In this part of the experiment, samples of 150 words (five paragraphs of 30 words each) were mounted on cardboard and presented to the readers in a controlled manner.

Results and Discussion

Data for the speed of reading comparisons are given in Table 1. Results for the control group (Group I) show that a "correction" must be made by adding 1.59 paragraphs to the mean for Form B. Examination of the results for Groups II and III reveals that the 8 point Opticon newsprint was read 0.92 of a paragraph more slowly than the book print, and that the 7 point Ionic newsprint was read more slowly than the book print by 1.01 paragraphs. These amount to a retardation in reading rate of 4.3 and 4.8 per cent respectively. The critical ratios in Column 10 of the table show that these differences are statistically significant.

These results demonstrate that commonly used newsprint even when printed in an optimum arrangement is read much more slowly than book print set in an optimum typographical arrangement.

The following factors probably operate to reduce the rate at which the newsprint was read: 1. The small size of newsprint type in comparison with the book type makes visual discrimination more difficult; 2. The lower brightness contrast between type and paper for the newsprint would adversely affect discrimination of the printed characters; and 3. Newspaper body types may not be as legible as book type faces. It is unlikely however that this third factor is important.

Results derived from reader opinions of relative legibility are given in Table 2. The order of judgments is 10 point book type ranked first,

⁸ Tinker, M. A., and Paterson, D. G. Studies of typographical factors influencing speed of reading. XIII. Methodological considerations. *J. appl. Psychol.*, 1936, 20, 132-145.

Table 1

Comparison of Speed of Reading Seven and Eight Point Newsprint with Ten Point Book Print

Differences given are for the mean score on Form A, 10 pt. book type, 20 pica line width, with 2 pt. leading, minus the mean score on Form B printed in newsprint as shown in comparison. Book type printed on eggshell and newsprint on newspaper stock. In each test group $N = 90$ college students.

Test Group	Comparison	Mean	P.E. Dist.	P.E. Mean	Diff. Between Means in			r	D. P.E. Diff.
					Para-graphs*	Per Cent	P.E. Diff.		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
I	A 10 pt. Cheltenham, 20 pica, 2 pt. leading	21.20	2.88	.30					
	B 10 pt. Cheltenham, 20 pica, 2 pt. leading	19.61	2.63	.28	0.00	.00	.00	.78	0.00
II	A 10 pt. Cheltenham, 20 pica, 2 pt. leading	21.58	3.13	.33					
	B 8 pt. Opticon, 12 pica, 1 pt. leading	19.07	2.73	.29	-0.92	4.27	.17	.86	5.37
III	A 10 pt. Cheltenham, 20 pica, 2 pt. leading	21.11	2.59	.27					
	B 7 pt. Ionic No. 5, 12 pica, 1 pt. leading	18.51	2.43	.25	-1.01	4.79	.17	.79	5.90

* The differences in column 6 are "corrected" by the amount of the difference between the mean scores of Form A and Form B of Test Group I which serves as a control group. The "correction" amounts to 1.59 paragraphs for each test group comparison.

Table 2

Book Type and Newsprint Ranked According to 117 Reader Opinions of Relative Legibility

Kind of Type	Average Rank	S.D.	Rank Order
10 Point Book Type	1.65	.79	1
8 Point Newsprint	1.68	.58	2
7 Point Newsprint	2.68	.60	3

followed by 8 point newsprint and then 7 point newsprint. Note, however, that there is actually very little difference in ranking 8 point newsprint and 10 point book print. As has been found before,⁹ judgments of legibility do not always agree with actual readability measurements.

Readers' opinions of pleasingness are listed in Table 3. The order

Table 3

Book Type and Newsprint Ranked According to 117 Reader Opinions of Pleasingness

Kind of Type	Average Rank	S.D.	Rank Order
10 Point Book Type	1.47	.66	1
8 Point Newsprint	1.70	.54	2
7 Point Newsprint	2.83	.46	3

from most to least pleasing is 10 point book type, 8 point newsprint and 7 point newsprint. Although there is some separation between mean ranks for the book type and the 8 point newsprint, the difference is not great. But the 7 point newsprint is considered definitely less pleasing than the others. As in the earlier report,¹⁰ pleasingness tends to agree with judged legibility.

Summary and Conclusions

1. The purpose of this investigation is to compare the readability of newsprint and book print.

2. Speed of reading 10 point Cheltenham book type was compared with speed of reading 8 point Opticon newsprint and with 7 point Ionic No. 5 newsprint.

3. Both kinds of newsprint were read significantly more slowly than the book print.

⁹ Tinker, M. A., and Paterson, D. G. Reader preference and typography. *J. appl. Psychol.*, 1942, 26, 38-40.

¹⁰ Tinker, M. A., and Paterson, D. G., *op. cit.*, 1942.

4. The slower rate of reading newsprint is apparently due to the greater difficulty of discriminating the printed characters in comparison with the book type which is larger and which involves greater brightness contrast between print and paper.

5. The 10 point book print and the 8 point newsprint are judged to be about equally legible, but the 7 point newsprint is considered to be far less legible.

6. The book print is judged to be most pleasing, the 8 point newsprint next most pleasing, and the 7 point newsprint least pleasing.

Received November 26, 1945.

Age of Starting to Contribute versus Total Creative Output

Harvey C. Lehman

Ohio University

In 1937, Professor E. T. Bell of the California Institute of Technology published this statement in his book, *Men of mathematics*:

"In 1902 and 1904 the Swiss mathematical periodical, *L'Enseignement Mathématique*, undertook an enquiry into the working habits of mathematicians. Questionnaires were issued to a number of mathematicians, of whom over a hundred replied. . . . To the question 'At what period . . . and under what circumstances did mathematics seize you?' 93 replies to the first part were received: 35 said before the age of ten; 43 said eleven to fifteen; 11 said sixteen to eighteen; 3 said nineteen to twenty; and the lone laggard said twenty-six" (1, p. 547).

The early interest in mathematics of those destined to become first-rank mathematicians has been asserted and commented on by numerous writers. This early interest is confirmed by the Swiss questionnaire study.

Some may doubt that the 93 mathematicians who responded to the first part of the above question would interpret it in exactly the same way. Others may suspect that some of the respondees were unable to recall precisely when they first became interested in mathematics. However, the problem of early interest can be approached from another angle as is done in the present study.

The present article presents information regarding the youngest chronological ages at which certain noted mathematicians made important contributions to their field. As here used, the words "important contributions" mean simply any contributions at all which are of sufficient merit to be cited and discussed by authorities in the history of mathematics.

Since the present writer is neither a mathematician nor a journalist he is unable to explain the mathematical contributions listed below any more clearly than the authors whose descriptions are quoted. Hence, the complete reliance in what follows on verbatim quotations.

If the reader finds this quoted material boresome, he can omit it without missing the main point of this discussion. These quotations serve a serious purpose, however, and some who may want to skip them at first will perhaps be motivated to read them after perusal of the main portion of this article.

The technical mathematical terms need not be understood in order to follow the gist of the quotations. In so far as the present study is concerned, the significance of this creative mathematical work lies in the fact that it was accomplished by mere boys not one of whom was over 21 years old at time of the indicated achievement.

Niels Henrik Abel, (1802–1829).

"Abel's first ambitious venture was an attack on the general equation of the fifth degree (the 'quintic'). All of his great predecessors in algebra had exhausted their efforts to produce a solution, without success. We can easily imagine Abel's exultation when he mistakenly imagined he had succeeded. . . . The supposed solution was of course no solution at all. This failure gave him a most salutary jolt; it jarred him onto the right track and caused him to doubt whether an algebraic solution was possible. He *proved the impossibility*. At the time he was about nineteen . . ." (1), p. 309f.).

Charles Babbage, (1792–1871).

"Charles Babbage invented a machine,¹ called a 'difference-engine,' about 1812. [Age 20]. Its construction was begun in 1822 and was continued for 20 years. The British Government contributed £17,000 and Babbage himself £6,000" (2), p. 485).

Charles Julien Brianchon, (1785–1864).

"His paper on curved surfaces of the second degree was published in the *Journal de l'École Polytechnique*, cahier 13, 1806. . . . This paper contains the famous theorem, known under the author's name, which together with Pascal's theorem [set forth at age 16] is at the very foundation of the projective theory of conic sections. . . . It is interesting to note that this article, which made the author's name familiar to every student of geometry, was written by him at the age of 21, while he was still in school" (3, p. 331).

Augustin-Louis Cauchy, (1789–1857).

"In February, 1811, Cauchy submitted his first memoir on the theory of polyhedra." [Age 21 years, 6 months] (1, p. 277).

Arthur Cayley, (1821–1895).

"His first work, published in 1841 when he was an undergraduate of twenty, grew out of his study of Lagrange and Laplace" (1, p. 381).

Alexis Claude Clairaut, (1713–1765).

"In 1731 was published his *Recherches sur les courbes à double courbure*, which he had ready for the press when he was sixteen. It was a work of remarkable elegance and secured his admission to the Academy of Sciences when still under legal age. In 1731 [age 18] he gave a proof of the theorem enunciated by I. Newton, that every cubic is a projection of one of five divergent parabolas" (2, p. 244).

William Kingdon Clifford, (1848–1879).

William Kingdon Clifford solved a problem in probability in 1866. [Age 18] (3, p. 540f.).

Leonhard Euler, (1707–1783).

"Euler's first independent work was done at the age of nineteen" (1, p. 144).

Jean-Baptiste-Joseph Fourier, (1768–1830).

"In December, 1789 Fourier (then twenty-one) went to Paris to present his researches on the solution of numerical equations before the academy.

¹ Probably an early form of the mechanical calculator.

This work advanced beyond Lagrange, and is still of value . . . it may be found in elementary texts on the theory of equations . . ." (1, p. 192).

Évariste Galois, (1811–1832).

" . . . Galois at the age of sixteen was already well started on his career of fundamental discovery . . ." (1, p. 366).

"Galois at seventeen was making discoveries of epochal significance in the theory of equations, discoveries whose consequences are not yet exhausted after more than a century" (1, p. 368).

"In Feb. 1830, at the age of nineteen . . . he composed three papers in which he broke new ground. These papers contain some of his great work on the theory of algebraic equations. It was far in advance of anything that had been done . . ." (1, p. 370).

During the night preceding the duel over a love affair which ended his life at age 20, Galois . . . "spent the fleeting hours feverishly dashing off his scientific last will and testament, writing against time to glean a few of the great things in his teeming mind before the death which he foresaw could overtake him. . . . What he wrote in those desperate last hours before the dawn will keep generations of mathematicians busy for hundreds of years. He had found, once and for all, the true solution of a riddle which had tormented mathematicians for centuries: under what conditions can an equation be solved? But this was only one thing of many" (1, p. 375).

Karl Friedrich Gauss, (1777–1855).

Gauss started his researches on pangeometry as early as 1792. [Age 15] (2, p. 304).

"He had already invented [at age 18] the method of 'least squares,' which today is indispensable in geodetic surveying, in the reduction of observations and indeed in all work where the 'most probable' value of anything that is measured is to be inferred from a large number of measurements" (1, p. 227).

"When not quite nineteen years old Gauss began jotting down in a copy-book very brief Latin memoranda of his mathematical discoveries. . . . Of the 146 entries, the first is dated March 30, 1796, [Age 18 years, 11 months] and refers to his discovery of a method of inscribing in a circle a regular polygon of seventeen sides. . . . He worked quite independently of his teachers, and while a student at Göttingen [Age 18 to 21] made several of his greatest discoveries. . . . The great law of quadratic reciprocity, given in the fourth section of Gauss' work, a law which involves the whole theory of quadratic residues, was discovered by him by induction before he was eighteen, and was proved by him one year later" (2, p. 435).

" . . . the entry for March 19, 1797, shows that Gauss had already discovered the double periodicity of certain elliptic functions. He was then not quite twenty. Again, a later entry shows that Gauss had recognized the double periodicity in the general case. This discovery of itself, had he published it, would have made him famous. But he never published it" (1, p. 229f.).

"At the age of twenty Gauss had overturned old theories and old methods in all branches of higher mathematics; but little pains did he take to publish his results, and thereby to establish his priority" (2, p. 434).

"Why did Gauss hold back the great things he discovered? . . . [A] statement which Gauss once made to a friend explains both his diary and his slowness in publication. He declared that such an overwhelming horde of new ideas stormed his mind before he was twenty that he could hardly control them and had time to record but a small fraction" (1, p. 227f.).

Wilhelm Jacob Storm van s'Gravesande, (1688–1742).

"His was another case of the early display of mathematical ability, his essay on perspective having attracted attention when he was only nineteen years old" (4, p. 526).

Edmund Halley, (1656–1742).

"... before he was twenty he communicated a paper to the Royal Society. So noteworthy had been his progress that in the very month in which he reached his twentieth birthday (November, 1676) he set out for St. Helena for the purpose of making astronomical observations. On the day before he was twenty-one he made the first complete observation of a transit of Mercury. So remarkable was his work at St. Helena that . . . the Royal Society elected him to a fellowship when he was only twenty-two" (4, p. 405).

William Rowan Hamilton, (1805–1865).

"... in his seventeenth year Hamilton had already begun his career of fundamental discovery. Before this he had brought himself to the attention of Dr. Brinkley, Professor of Astronomy at Dublin, by the detection of an error in Laplace's attempted proof of the parallelogram of forces" (1, p. 343).

"At the age of twenty-three he published the completion of the 'curious discoveries' he had made as a boy of seventeen, Part I of *A theory of systems of rays*, the great classic which does for optics what Lagrange's *Mécanique analytique* does for mechanics and which, in Hamilton's own hands, was to be extended to dynamics, putting that fundamental science in what is perhaps its ultimate, perfect form" (1, p. 346).

Charles Hermite, (1822–1901).

"The *Nouvelles Annales de Mathématiques*, a journal devoted to the interests of students in the higher schools, was founded in 1842. The first volume contains two papers composed by Hermite while he was still a student at Louis-le-Grand. [Age about 20]. The first is a simple exercise in the analytic geometry of conic sections and betrays no originality. The second, which fills only six and a half pages in Hermite's collected works, is a horse of quite a different color" (1, p. 450).

Ernst Eduard Kummer, (1810–1893).

"In his third year at the University Kummer solved a prize problem in mathematics and was awarded his Ph.D. degree at the age of twenty-one" (1, p. 512).

Joseph Louis Lagrange, (1736–1813).

"At the early age of nineteen he sent a solution of the isoperimetrical problem to Euler, in which he announced the principle of the calculus of variations.

"This memoir inaugurated a new period in the calculus of variations and was esteemed very highly by Euler, who observed that the methods of Lagrange were more general than his own" (5, p. 240f.).

Guillaume François Antoine de L'Hospital, (1661–1704).

"When only fifteen he was one day at the Duc de Roanne's and heard some mathematicians speaking of a difficult problem of Pascal's. To their surprise he said that he thought he could solve it, and in a few days succeeded" (4, p. 384).

Colin Maclaurin, (1698–1746).

At the age of twenty-one he took to his London printer "his *Geometria Organica*, containing a new and remarkable mode of generating conics, known by his name . . ." (2, p. 228).

James Clerk Maxwell, (1831–1879).

"At the age of fifteen he published a paper on oval curves" (6, p. 251).

Gaspard Monge, (1746–1818).

At about age sixteen Gaspard Monge originated descriptive geometry. He "was at once given a minor teaching position to instruct the future military

engineers in the new method. Monge was sworn not to divulge his method, and for fifteen years it was a jealously guarded military secret" (1, p. 185).

François Nicole, (1683–1758).

"He was a boy of unusual promise, having shown his genius in geometry by rectifying the cycloid at the age of nineteen" (4, p. 472).

Blaise Pascal, (1632–1662).

"Before the age of sixteen (about 1639)² he had proved one of the most beautiful theorems in the whole range of geometry" (1, p. 76).

"Presently the family received a somewhat formal visit from Descartes. He and Pascal talked over many things, including the barometer. There was little love lost between the two. For one thing, Descartes had openly refused to believe the famous *Essai pour les coniques* had been written by a boy of sixteen" (1, p. 80).

"At the age of nineteen he invented a computing machine that served as a starting point in the development of the mechanical calculation that has become so important in our time. That he should have been permitted to present one of these machines to the king and one to the royal chancellor shows the esteem in which he must have been held" (4, p. 382).

Simeon Denis Poisson, (1781–1840).

"At eighteen he wrote a memoir on finite differences which was printed on the recommendation of A. M. Legendre" (2, p. 466).

Georg Friedrich Bernhard Riemann, (1828–1866).

"According to Dedekind, 'Riemann recognized in . . . partial differential equations the essential definition of an [analytic] function of a complex variable. Probably these ideas, of the highest importance for his future career, were worked out by him in the fall vacation of 1847 [Riemann was then twenty-one] for the first time" (1, p. 489f.).

James Joseph Sylvester, (1814–1897).

"About the age of 16 he was awarded a prize of \$500.00 for solving a question in arrangements for contractors of lotteries in the United States" (2, p. 343).

William Thomson (Lord Kelvin), (1824–1907).

Rediscovered independently, in 1845, the principle of inversion called by Liouville the transformation by reciprocal radii. [Age 21] (2, p. 292).

The foregoing quotations permit glimpses of the mathematical maturity that has been attained by some individuals before they were 22 years old. These quoted statements should help readers to take seriously what follows. Standing alone, however, they are inadequate for our present purpose. As is indicated by its title, this study concerns itself with the relationship between age of starting to contribute creative work and total creative life output. We want to know, among other things, whether these youthful contributors fulfilled the promise of their early youth. The complete picture must include both the answer to this query and also information regarding persons who have started to make their intellectual contributions at each successive older age level. A bird's-

² Authorities differ on Pascal's age when this work was done, the estimates varying from fifteen to seventeen.

eye view of our major findings with reference thereto can best be described by means of scattergrams constructed in a manner that will now be described.

Figure 1 presents: (1) the chronological age at which each of 306 deceased chemists made his first chemistry contribution of sufficient merit to be included in T. P. Hilditch's *A concise history of chemistry* (7), versus (2) the total number of contributions by each chemist included

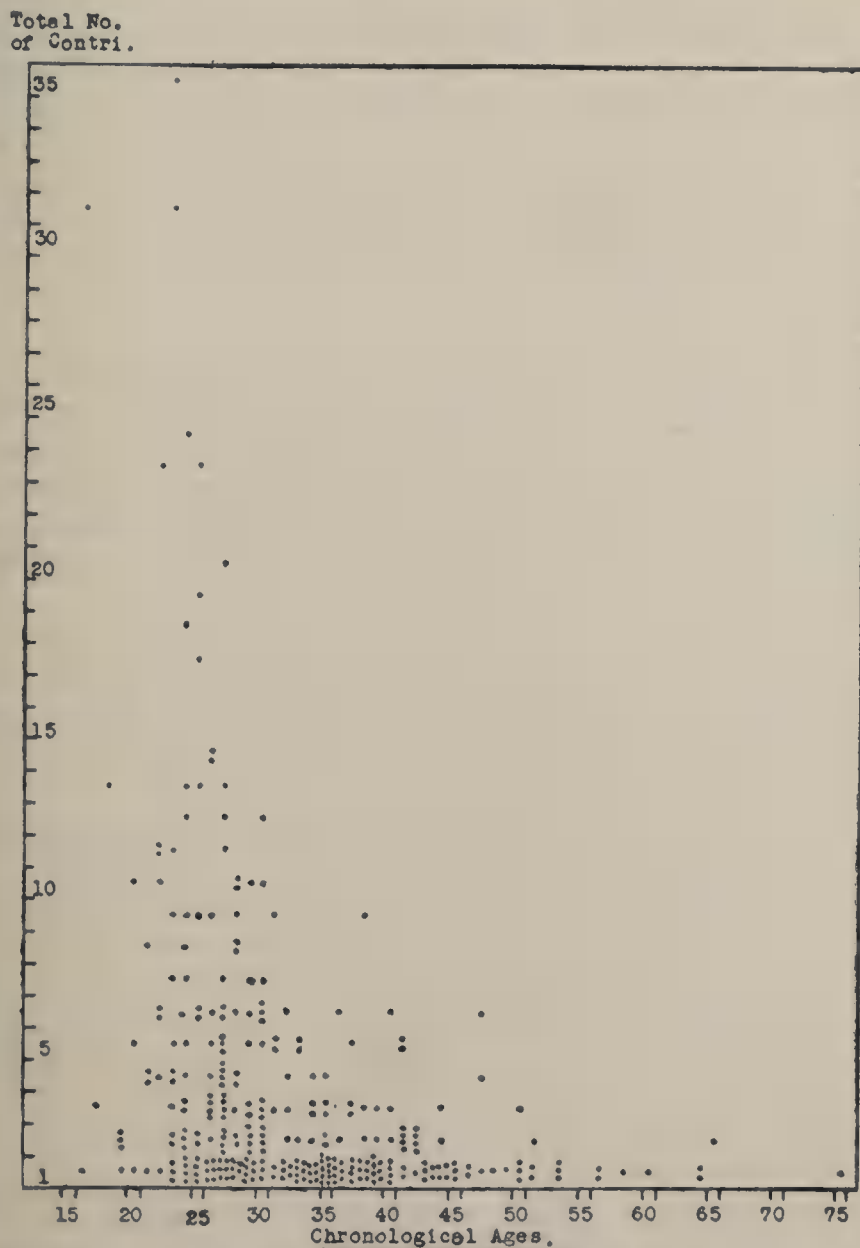


FIG. 1. The age at which each of 306 chemists made his first important chemistry contribution versus the total number made by each.

therein. For example, this figure reveals that Justus von Liebig, who is credited with 31 chemistry contributions in Hilditch's history, made the first of those 31 contributions at age 16. Figure 1 shows also that the one person who made the largest number of contributions cited by Hilditch, Emil Fischer with 35 contributions, did his first important research in chemistry at age 23.

Figure 1 reveals in general that the larger the total number of notable chemistry contributions made by a given individual, the younger the chronological age at which that individual started his research career. Thus, of the 10 most prolific contributors cited by Hilditch (see upper left-half of Fig. 1), only one made his initial contribution as late as age 27. Of the 29 largest contributors pictured in Figure 1, only 2 made their first important contribution as late as age 20. And of the 45 heaviest contributors, one only made his first contribution later than age 31. On the whole, it seems apparent that the later one starts to achieve in chemistry, beyond the early twenties, the smaller the probable total of one's outstanding research contributions.

Near the base-line of Figure 1, it may be seen that of 140 chemists each of whom made one contribution only that is described by Hilditch, one individual made his first and only notable chemistry contribution as late as age 75. It seems logical to infer that, although there is no deadline beyond which it is impossible to make one's initial contribution, age 75 is too old to start contributing if one hopes to make more than one important contribution.

Of 45 persons each of whom made two significant contributions to chemistry, only 1 of them did his first research as late as age 65, and 43 of the 45 did their first notable research prior to age 45. Similarly, of 31 persons each of whom made 3 significant contributions, only 2 of them did their initial outstanding work as late as age 40. Finally, of 16 persons, each of whom made but 4 contributions, only 1 of them did his first important research work beyond age 35.

Oliver Wendell Holmes, Jr., is said by several of his biographers to have held the belief or superstition that if genius is to be exhibited at all it must be displayed prior to age 40. It is said that Holmes once remarked: "If you haven't cut your name on the door of fame by the time you've reached 40, you might just as well put up your jackknife."

Figure 1 suggests that, in so far as creative chemistry is concerned, Holmes's foregoing figure of speech should perhaps be revised to read as follows. If an individual is to make a chemistry contribution which ranks in importance with those cited in Hilditch's history, the chances are 6 in 7 that the individual will have completed his first important re-research before he has passed age 40. This modification of Holmes's

statement seems justified in view of the finding that of the 306 chemists for whom age data are presented in Figure 1, not less than 83% did their first important chemistry research prior to age 40.

The words "not less than 83%" were employed in the above because the time lag between date of accomplishment and date of announcement thereof is not always known. It seems likely that, if the full story were known, some of the other 17% may also have done their initial research at younger ages than the available record reveals.

The data that are set forth graphically in Figure 2 were obtained from Cajori's *A history of mathematics* (2). This figure presents: (1) the age at which each of 444 deceased mathematicians made his first contribution of sufficient merit to be mentioned and discussed in Cajori's history, versus (2) the total number of contributions by each mathematician discussed therein for whom age data were available. Mathematicians will be interested to know that the correlation ratio between these two variables is $-.61$.

Figure 2, like Figure 1, suggests that there is a lower age limit or threshold, prior to which important mathematical contributions are not likely to be made. This lower age limit occurs at a younger age level, however, than might have been anticipated by many, namely, somewhere in the late teens or early twenties, the exact lower limit depending upon the type, and perhaps even more upon the quality of the contribution that is under consideration.

Figure 3 is based upon data obtained from 10 histories of physics. It reveals: (1) the age at which each of 388 deceased outstanding physicists made his first contribution deemed worthy of citation and discussion by one or more of the 10 historians, versus (2) the total number of contributions that was made by each of the 388 physicists. Comments already made with reference to Figures 1 and 2 should enable readers to interpret Figure 3 without further aid.

Table 1 presents statistical information regarding those who made many versus those who made few important contributions to chemistry. This table reveals that those who made larger numbers of notable chemistry contributions did their first work at a younger average age (column 3) and their final work at an older average age (column 4) than did those who made fewer outstanding contributions. In other words, the more prolific contributors of outstanding research got "on the beam" earlier and they stayed on longer.

Tables similar to Table 1 were constructed for 20 different kinds of creative endeavor. Like Table 1, the latter all reveal that, as compared with the minor contributors, the major contributors to a given field accomplished their first important research at younger average ages and their last important work at older average ages.

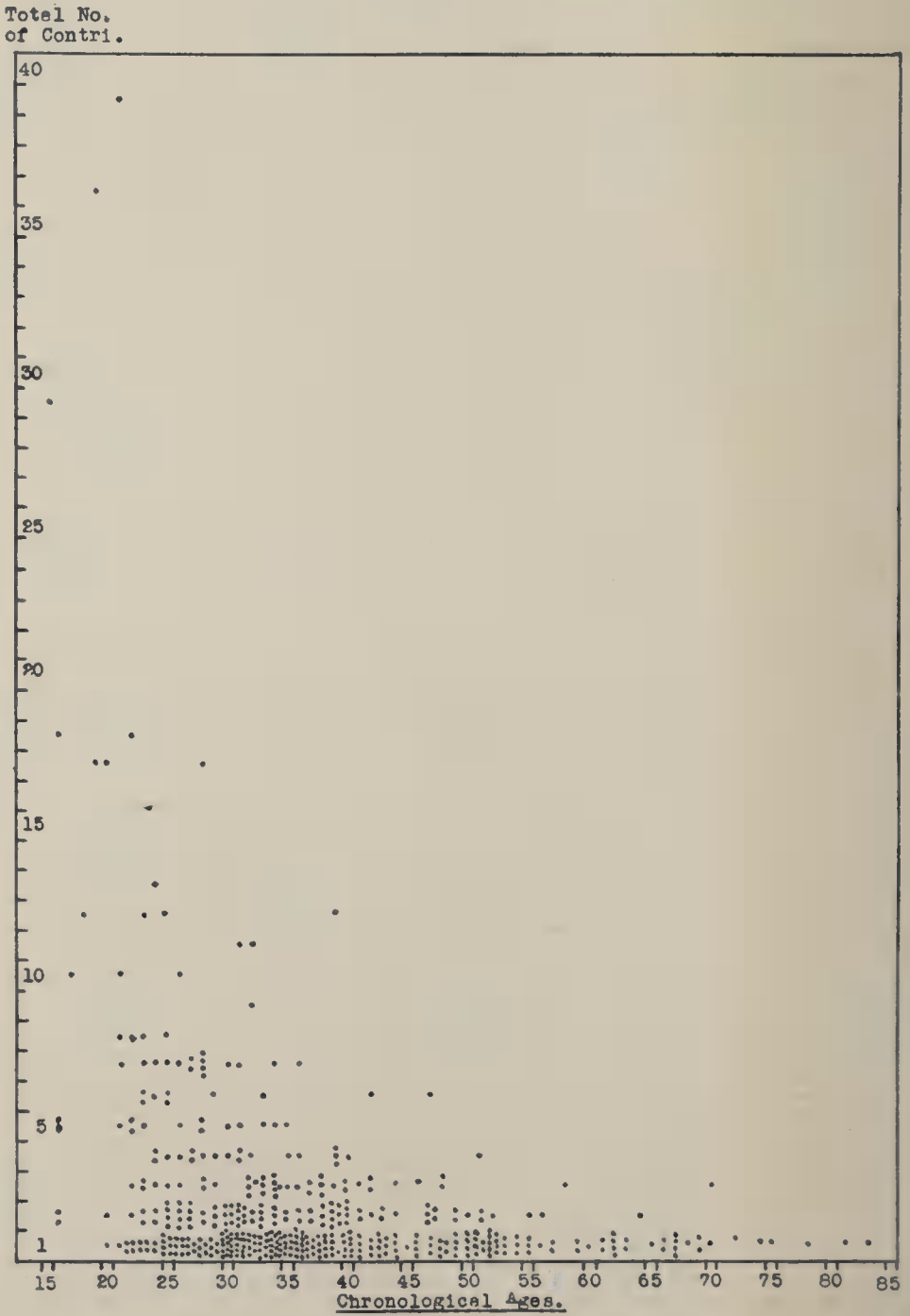


FIG. 2. The age at which each of 444 mathematicians made his first important contribution to mathematics versus the total number made by each.

Scattergrams, like Figure 1, were likewise constructed for 20 types of creative work. Space limitations preclude publication of the 20 scattergrams and the data from which they were made.

For 10 fields of work comparisons were made between the average total life output of individuals who made their first contributions at ages

Total No.
of Contri.

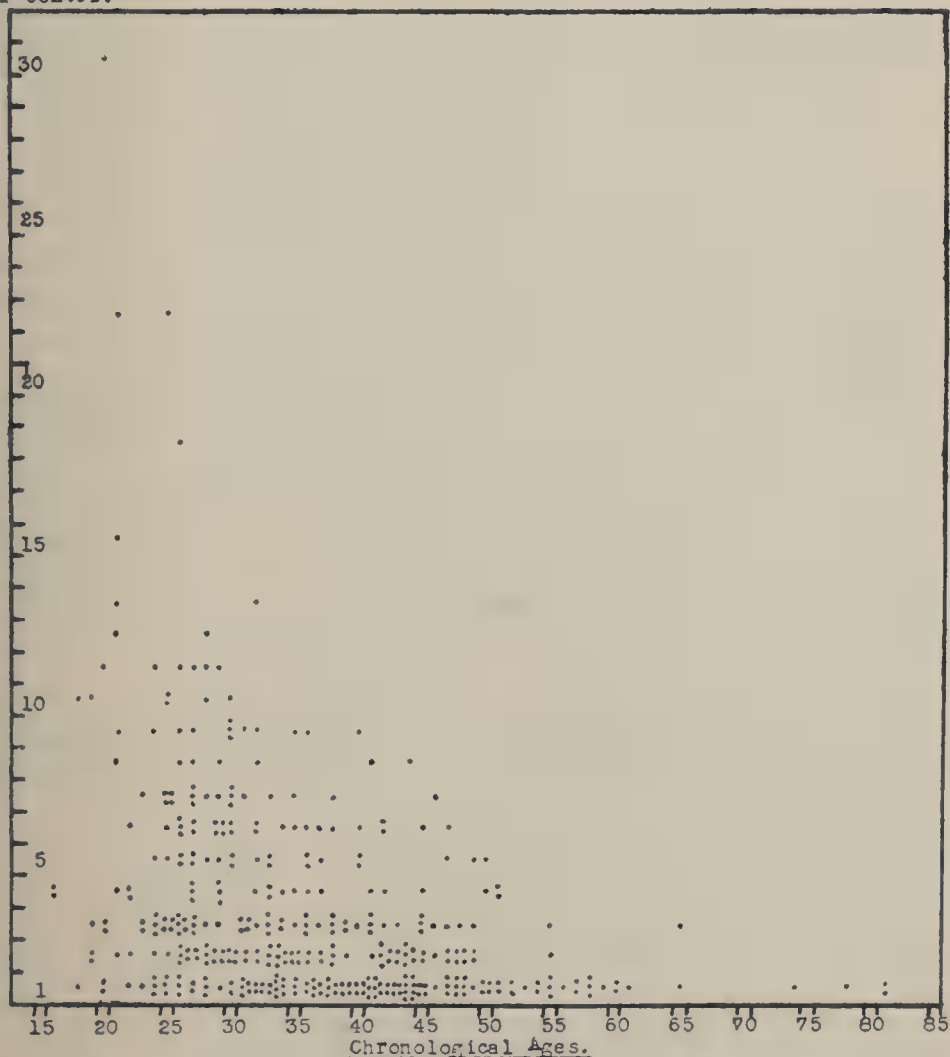


FIG. 3. The age at which each of 388 physicists made his first important contribution to physics versus the total number made by each.

15 to 19 inclusive, versus the average total output of others who started making their contributions to the same fields of endeavor from ages 20 to 24 inclusive. Both the mean and the median output of those groups which started contributing at the younger age interval were 24% greater,

Table 1
More Prolific versus Less Prolific Contributors to Chemistry

No. of Men	No. of Contri. Cited by Hilditch	Average Age at Time of First Contri.	Average Age at Time of Last Contri.	Average No. of Years Between First and Last Contri.
135	1 only	36.27	36.27	0
241	1-5	34.10	38.59	4.49
38	6-10	27.95	48.05	20.10
13	11-15	24.69	47.92	23.23
4	16-20	25.25	50.75	25.50
6	21 or more	22.17	60.17	38.00

Sufficient data were available in 20 fields of endeavor to permit trustworthy comparisons between the average output of groups that started to contribute at ages 20 to 24, versus other groups that started contributing from ages 25 to 29 inclusive. Both the mean and the median output were 19% greater for those groups which started contributing when younger.

The fact that delay in starting to contribute at almost any given age level is still likely to make a difference in total average output may be seen by inspection of the scattergrams here presented. The approximate amounts of the decrements up to age 50 are given, by 5-year intervals, in the following tabulation.

Mean and median percentages of decrease in average total life output by groups which started their contributions at successive 5-year intervals. In each instance the comparison is made with the groups which started to contribute during the preceding 5-year interval.

Age Intervals.....	20-24	25-29	30-34	35-39	40-44	45-49
No. of Groups.....	10	20	20	20	20	20
Mean decrease.....	24%	19%	17%	14%	10%	14%
Median decrease.....	24%	19%	24%	19.5%	12.5%	14.5%

Table 2 lists the 16 English authors who have the largest number of works listed in Ryland's *Chronological outlines of English literature* (8). This table reveals that for these 16 very prolific contributors, the mean age at time of making their first contribution was 23.3 years; the mean age at time of making their last contribution was 58.0 years; and the mean interval between the first and the last contributions was approximately 35 years.

One need not be an advanced student of English literature to realize that it would be difficult and perhaps impossible to find any other list of 16 English authors as distinguished as those whose names are found in Table 2. Those who believe that the renowned author attains his pre-eminence without effort should examine carefully the number of contributions made by each writer whose name appears in Table 2.

The names of 152 of Germany's best-known, and probably most able, literary men were obtained from a highly select bibliography (9) which lists an average of only 3.0 works per author. The date on which each of

Table 2
Sixteen Most Prolific Contributors to English Literature

Individual	No. of Contri. Cited by Ryland	Age at Time of First Contri.	Age at Time of Last Contri.	No. of Years Between First and Last Contri.
Shakespeare.....	42	24	49	25
Sir Walter Scott.....	42	25	60	35
Dryden.....	40	28	74	46
Lord Bulwer-Lytton.....	39	24	70	46
Fletcher.....	36	31	46	15
Byron.....	32	19	36	17
Robert Browning.....	32	21	77	56
Pope.....	32	21	50	29
Dickens.....	30	22	58	36
Swift.....	30	31	71	40
Ruskin.....	28	20	66	46
Fielding.....	28	21	47	26
Tennyson.....	25	18	80	62
Shelley.....	25	18	30	12
Swinburne.....	25	24	52	28
Defoe.....	25	26	67	41
Averages.....	31.9	23.3	58.0	34.8

these authors published his first work was then found in Kosch (10) who seems to have listed almost everything that these 152 authors published, namely, an aggregate of 2,935 works, the average number of works per author being 19.3. It is significant that 35% of the 152 distinguished German authors started to publish at age 24 or younger, and that 11% of them started to publish while still in their teens.

A second group of 473 less distinguished German writers, listed by Kosch only, was next obtained by canvassing Kosch's *Lexikon* from A to Ca, inclusive. The average output for this less distinguished group was

only 3.4 works per author. This is less than 20% of the average output of the more distinguished Germans.

The less distinguished group of German authors is also characterized by a much smaller per cent of early starters. In contrast to the more select group of 152 German writers listed by Priest, of whom 35% started

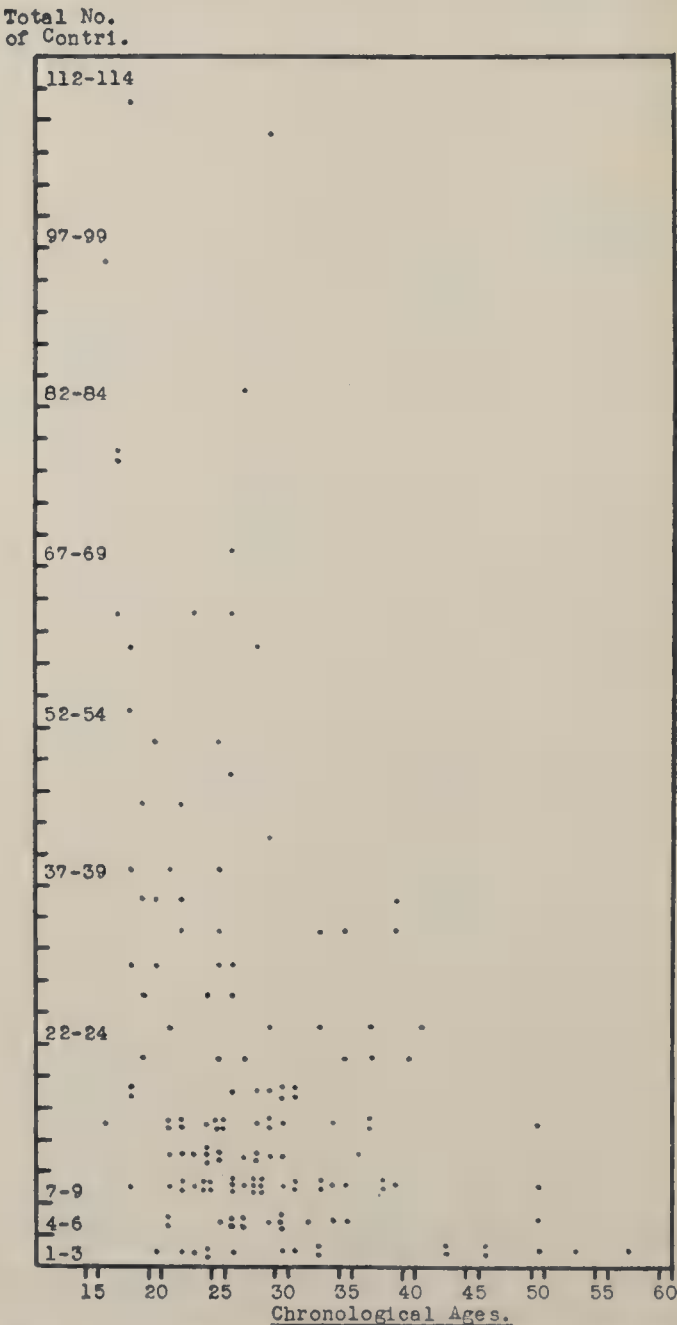


FIG. 4. The age at which each of 152 highly distinguished German authors published his first work versus the total number of works published by each.

to publish at age 24 or younger, only 9% of the less distinguished German authors started to publish when as young as age 24 or less. (See Table 3.)

Figure 4 presents: (1) the age at which each of the 152 more eminent German writers published his first work, versus (2) the total number of contributions made by each. Figure 5 sets forth similarly: (1) the age at which each of 152 less distinguished German authors, listed by Kosch

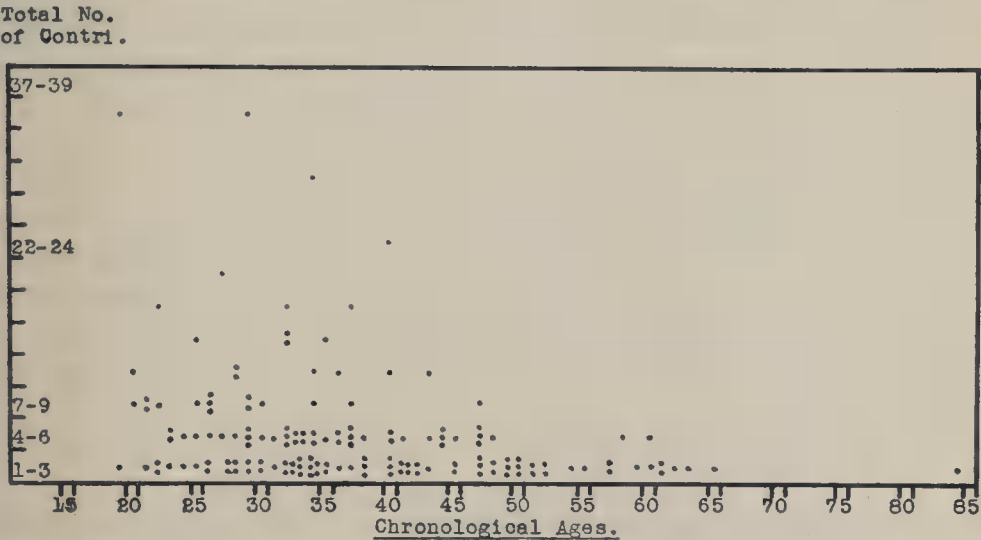


FIG. 5. The age at which each of 152 less distinguished German authors published his first work versus the total number of works published by each.

only, published his first work, versus (2) the total number of publications of each. In order to make Figures 4 and 5 more directly comparable, data for only 152 of the 473 less distinguished Germans, taken alphabetically, are set forth in Figure 5.

Table 3 makes possible a comparison of the average productiveness of the youthful German starters who later were identified by Priest as

Table 3

Age Data Regarding More Distinguished versus Less Distinguished German Authors

Main Group	No. of Indiv.	Starters at Age 24 or Less		
		Per cent of Main Group	Total Works Listed by Kosch	Per cent of Main Group's Total Output
152 authors listed by both Priest and Kosch	53	35%	1,598	45%
473 less distinguished authors listed by Kosch only	43	9%	440	28%

highly distinguished, versus the productiveness of other youthful starters who failed to achieve such distinction. Notice in Table 3 that the 53 more distinguished German authors who started contributing at age 24 or younger produced a total of 1,598 works, the average being 30.2 works per author! This high average is almost the same as that of the 16 prolific English authors listed in Table 2.

The 43 other Germans who started to publish when equally young but who are not listed by Priest, published an average of slightly more than 10 works each. Their average output is thus only about a third as great as is the average of the youthful starters who attained greater eminence. Their average output, nevertheless, is about three times as great as is the average of others in the less distinguished group who started to publish when beyond age 24.

By way of summary it may be said that: (1) as compared with the less distinguished, the more distinguished group of German authors was far more prolific, (2) a larger percentage of the more distinguished group started to publish during their late teens and early twenties, and (3) within each group the more youthful starters exhibited greatest productivity.

Since similar results were found when numerous other such comparisons were made, it seems fair to conclude that, as compared with the average individual, our most distinguished creative thinkers have usually possessed, among other things, an astonishing capacity for hard patient work.

For example, with reference to three of the mathematicians, whose initial work is described by use of quotations in the forepart of this article, one authority in the history of mathematics remarks: "For prolific inventiveness Euler, Cauchy, and Cayley are in a class by themselves . . ." (1, p. 378). More specifically with regard to the industry of Leonhard Euler the same historian asserts:

"The extent of Euler's work was not accurately known even in 1936, but it has been estimated that sixty to eighty quarto volumes will be required for the publication of his collected works" (1, p. 139).

Of Cauchy he present the following details:

"During the last nineteen years of his life he produced over 500 papers on all branches of mathematics, including mechanics, physics, and astronomy. Many of these works were long treatises" (1, p. 289). "His total output is 789 papers (many of them very extensive works) filling twenty-four large quarto volumes" (1, p. 292).

With reference to Cayley we find:

". . . his massive Collected Mathematical Papers (thirteen large quarto volumes of about 600 pages each, comprising 966 papers) will suggest profitable forays to adventurous mathematicians for generations to come" (1, p. 402).

After reading the above statements, one can hardly refrain from wondering whether Euler, Cauchy, Cayley, and the other prodigious workers, of whose industry fleeting glimpses may be obtained herein, were not criticized by some of their contemporaries for overproduction.

A list of notable oil paintings was obtained by making a composite study of 60 different books which contain lists of so-called master paintings. This procedure utilizes the collective judgments of art critics and historians who have published evaluations under their own signatures and who must, therefore, have tried conscientiously to make sound evaluations.

In what follows it is assumed: (1) that this large number of independent critics have exhibited no constant prejudice for or against any one particular age group, and (2) that careful study of the frequency with which paintings have been listed by the various compilers should enable one to identify the really great paintings.

The next step in the analysis of the data was to omit all paintings which were found to be listed only once or twice in the 60 compilations on the theory that this procedure would tend to eliminate eccentric judgments. That is to say, if a given painting were judged to be a great artistic work by only 1 or 2 or the 60 compilers, it was assumed that that particular painting would be less likely to possess genuine merit than would another painting which had been chosen by 3 or more of the 60 compilers. Those paintings which were chosen by 3 or more compilers will be referred to hereinafter as "superior" paintings.

When the data for artists who produced superior paintings were partitioned in various ways and analyzed, it was found that the more recently-born artists had achieved an average of only 2.93 superior paintings per individual, whereas, the earlier-born artists had achieved an average of 4.55 superior paintings.

Further scrutiny revealed also that not less than 35% of the earlier-born, and more distinguished, artists did their first oil painting (not necessarily a superior one) at age 24 or younger, and that not less than 18% of them did their first painting at age 19 or less. (See Figure 6.) In contrast to the more distinguished group, the more recently-born artists, who seem to have achieved a somewhat less enviable record than the earlier-born, are characterized also by a smaller percentage of very youthful starters, only 15% of the more recently-born group having started their painting career at age 24 or younger. (See Table 4.)

Table 5, which presents data for composers of grand opera, was constructed in a manner similar to Table 4. Table 5 reveals once again that the more distinguished of two groups of creative workers includes a larger percentage of talented youthful starters.

Total No.
of Contri.

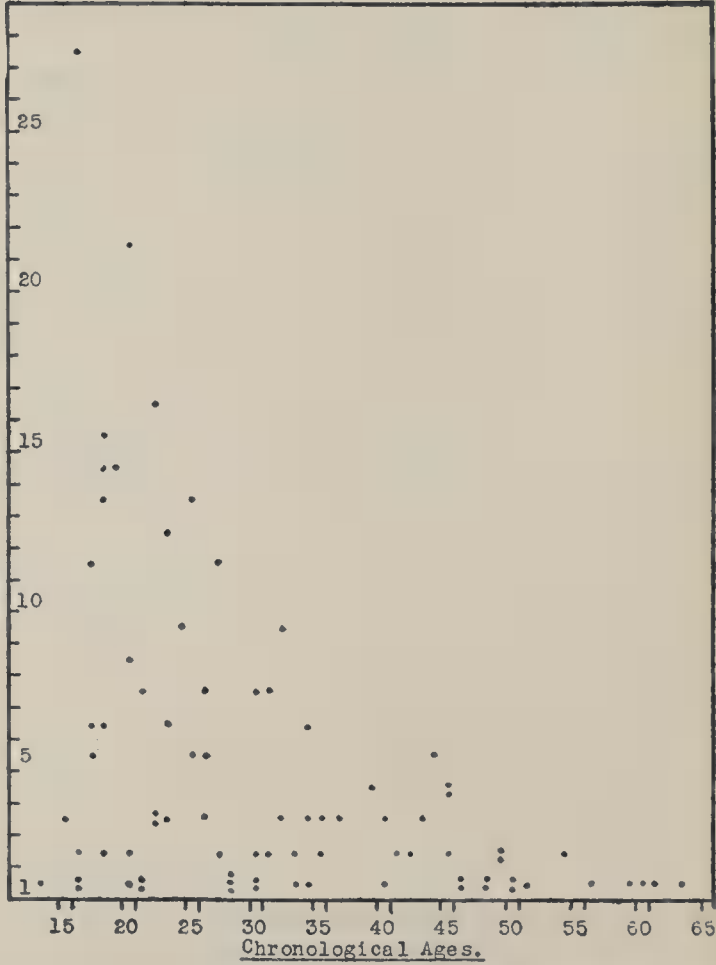


FIG. 6. The age at which each of 80 very distinguished artists did his first painting in oil versus the total number of superior pictures produced by each.

Table 4

Age Data Regarding More Distinguished versus Less Distinguished Painters in Oil

Main Group	Starters at Age 24 or Less			
	No. of Artists	Per cent of Main Group	Total No. Superior Paintings	Per cent of Main Group's Total Output
80 artists born prior to 1630 who achieved an average of 4.55 superior paintings.	29	36%	214	57%
80 artists born from 1630 to 1850 who achieved an average of only 2.93 superior paintings.	12	15%	47	20%

Table 5

Age Data Regarding More Distinguished versus Less Distinguished
Composers of Grand Opera

Main Group	Starters at Age 24 or Less			
	No. of Men	Per cent of Main Group	Total No. of Operas	Per cent of Main Group's Total Output
88 men who composed 191 superior grand operas each of which appeared 3 or more times on a composite list	32	36%	80*	42%*
559 other men who produced 1,723 operas of lesser merit†	82	15%	451†	26%†

* Superior operas only.

† Operas listed by Pratt, W. S.: *The new encyclopedia of music and musicians*. New York: The Macmillan Co., 1924. Pp. vi + 967.

When scattergrams for quite different kinds of creative achievement are based upon works of only very superior quality, they exhibit much similarity. For example, Figure 7 presents: (1) the ages at which each of 207 noted French and German philosophers published his first work (not necessarily a very superior one), versus (2) the total number of superior philosophical works by each, i.e., works which were cited and discussed in 3 or more of 50 histories of philosophy.

Although the data for these different types of creative work are sufficiently alike to support our main conclusions, one cannot judge accurately, by mere inspection of our tables and graphs, which kind of creative work comes earliest or which can be continued longest. This is because the several kinds of creativity cited herein have not been equated as regards their quality or merit. For example, the first and the last chemistry contributions may be somewhat more select (or somewhat less select) than are, let us say, the first and the last philosophical contributions. It is true also that information regarding first and last contributions may not be equally available in every field. Unequal amounts of time lag between date of achieving and date of announcement thereof are also a possibility. For all of these reasons, the data herein should be regarded as approximate only, not as mathematically exact or directly comparable.

As a supplement, the following literary exposition is perhaps *à propos* since the concluding statement is so abundantly validated by our statistics.

Total No.
of Contri.

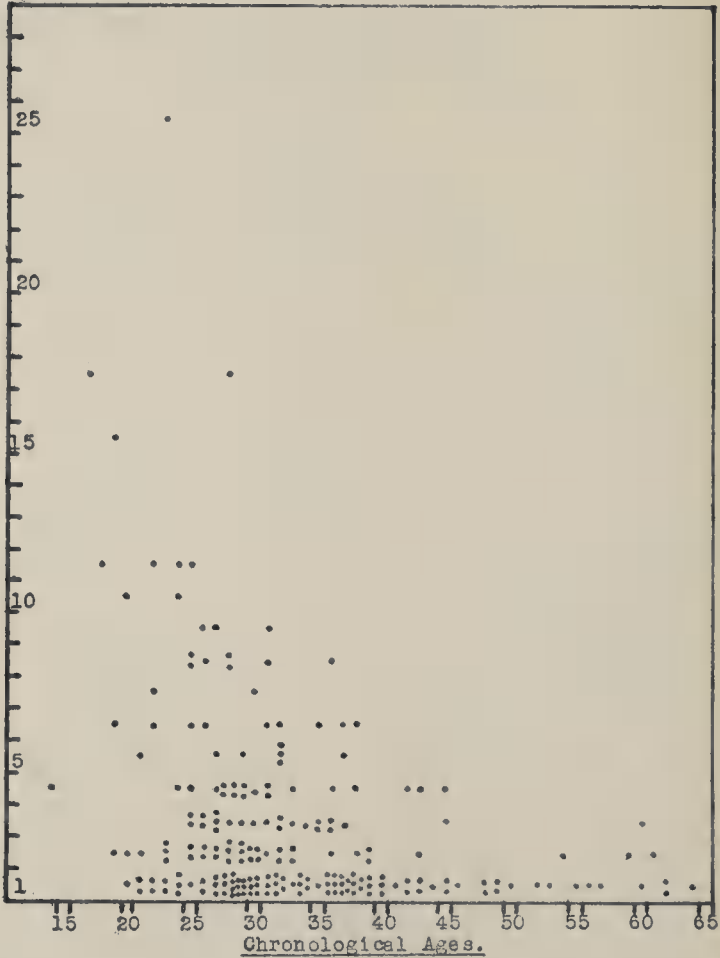


FIG. 7. The age at which each of 207 French and German philosophers published his first work versus the total number of superior works produced by each.

"When Lindbergh flew to France—at just 25—every newspaper had to dwell upon his youth. He was a mere kid. Yet he was as old as Keats was at death. He was a year older than Pitt was when he became prime minister of England. He was eight years older than Mendelssohn was when he composed his overture to *A midsummer night's dream*. John Ericsson, who did many things besides build the *Monitor*, was a draftsman at 12 and a full-fledged engineer at 15. Chatterton finished at 18; Galois, the mathematician, at 20. Jane Austin was writing one of her best novels [*Pride and prejudice*] at 21. . . .

"Anyone can leaf through a dictionary of biography and make similar lists in a half hour. In other words, much of the significant record of the human race has been made by men and women scarcely older than the hundreds of thousands of students who mull along in crowd fashion, year after year, in our undergraduate colleges" (11, p. 16).

This discussion will end with the observation that all of the renowned benefactors of humanity in the following list were less than 25 years old

(Lindbergh's age when he flew to France) at the time they did their first important creative work.

- 56 of the chemists as shown in Figure 1.
- 53 mathematicians (see Figure 2).
- 41 physicists (see Figure 3).
- 53 distinguished German authors (see Figure 4 and Table 3).
- 11 (or 69%) of the 16 English authors listed in Table 2. (For 520 less distinguished English authors listed by Ryland, who produced an average of only 3.98 works per author, the corresponding figure is 15%).
- 29 noted painters in oil (see Figure 6 and Table 4).
- 30 eminent French and German philosophers (see Figure 7).
- 32 composers of superior grand operas (see Table 5).

There is reason to suspect that the above sample findings, not one of which is exhaustive, can be duplicated easily in almost every creative field. On the whole, our findings suggest that those destined to go far have started early and moved rapidly.

Although we know that the earlier starters produced on the average both more and better creative work than did those who started to contribute later in life, it would be even more valuable if we could ascertain whether the earlier start tends to make one's best creative work, even better than it otherwise would be. It may be argued, with some justice perhaps, that it was the most brilliant individuals who started making their contributions earliest, and that both the quantity and the quality of their output were due solely, or chiefly, to their unequalled brilliance rather than to their early start. Since the phenomena here considered cannot be subjected to rigorous controlled experimentation, one can only speculate with reference to this latter possibility.

Received October 19, 1945.

References

1. Bell, E. T. *Men of mathematics*. New York: Simon and Schuster, 1937. Pp. xxi + 592.
2. Cajori, F. *A history of mathematics*. New York: The Macmillan Co., 1922. Pp. viii + 516.
3. Smith, D. E. *A source book in mathematics*. New York: McGraw-Hill Book Co., 1929. Pp. xvii + 701.
4. Smith, D. E. *History of mathematics*. Vol. I. *General survey of the history of elementary mathematics*. Boston: Ginn and Co., 1923. Pp. xxii + 596.
5. Miller, G. A. *Historical introduction to mathematical literature*. New York: The Macmillan Co., 1927. Pp. xiii + 302.
6. Cajori, F. *A history of physics in the elementary branches*. New York: The Macmillan Co., 1899. Pp. viii + 322.

7. Hilditch, T. P. *A concise history of chemistry*. New York: D. Van Nostrand Co., 1911. Pp. ix + 263.
8. Ryland, F. *Chronological outlines of English literature*. London: Macmillan & Co., 1910. Pp. xii + 351.
9. Priest, G. M. *A brief history of German literature*. New York: Charles Scribner's Sons, 1909. Pp. xii + 366.
10. Kosch, W. *Deutsches Literature-Lexikon: Biographisches und Bibliographisches Handbuch*. In zwei Bände. Halle: (Saale), Max Niemeyer Verlag. 1927.
11. Brown, R. Youth in the saddle. *The Reader's Digest*, Mar. 1934, 24, p. 17.

The Use of the Harrower-Erickson Multiple Choice Rorschach Test with a Selected Group of Women in Military Service

Capt. Marjorie Case Winfield, USMCR *

When the plan to send members of the Marine Corps Women's Reserve overseas was authorized a staging area was established where all personnel were sent for a period of about two weeks immediately prior to being shipped overseas, for the purpose of medical and dental examinations, outfitting and other processing, a minimum amount of re-training and final screening.

Since the tour of duty was to be a minimum of two years, with no state side furlough or leave during that time, and no guarantee of immediate return upon completion of the two years, it was important to send women Marines who could stand the monotony as well as the excitement, the limitations and frustrations as well as the adventures of such an assignment. Consequently, the qualifications for duty were set high. Anyone desiring an overseas assignment had to volunteer for it, and to be eligible a woman Marine had to have had a minimum of six months active duty in the Marine Corps, exclusive of recruit or specialist training. She must have had a good conduct record, a good health record with no misconduct status, no courts martial, and a good work record. She must have been recommended by her Commanding Officer and she must have "demonstrated in her military service a sense of responsibility, maturity, adaptability and emotional stability."

These requirements automatically excluded the obviously misfitted, those with records of previous maladjustments of one sort or another. However, there was still the possibility that certain members of this selected group, under strange and more complicated circumstances, would become problem cases. To aid in the detection of potential misfits two tests, the Harrower-Erickson Multiple Choice Test (for use with Rorschach Cards or Slides) ¹ and the Minnesota Multiphasic Personality Inventory, (MMPI), were given to the first group of 181 enlisted women the day after their arrival in the staging area.

* This paper is not to be construed as the official opinion or conclusions of the USMC.

¹ Henceforth in this paper, in order to avoid confusion with the Harrower-Erickson Group Rorschach Test, this test will be referred to as the Multiple Choice Rorschach or abbreviated MCR.

The sole purpose in giving these two tests to the group was to give the officers charged with the responsibility of the final selection some concrete indication of those individuals least likely to maintain a satisfactory adjustment. It was intended that the women so indicated by the tests would be closely observed by the staging area staff during the two week staging period which, of necessity, was one of strain and tension. Neither test was intended to be used for differential diagnosis nor were they to be considered either separately or combined as sufficient reason to exclude a woman from going overseas, unless substantiated by other evidence obtained through careful observation.

Administration and Scoring

The MCR was given to groups of 35 enlisted women at a time. Instructions for marking the blanks were given as they are written on the cover page of the test form. Each Rorschach slide was thrown on the screen for a period of 1½ minutes with the lights off, and for an equal amount of time with the lights on, making a total of 3 minutes for each slide.

The MMPI was given after the MCR usually with about a half hour interim, with the instructions given as they are on the cover page of the booklet.² Scoring was done by machine.

Prior to establishing the testing program the scoring method of the MCR was discussed with the author of the test whose opinion it was that the circumstances might warrant penalizing the individuals who checked "Nothing At All" or who made no markings whatsoever in one of the A, B, or C sections, both of these answers indicating a failure. As a consequence of this discussion, two methods of scoring were used on all blanks:

- (a) Each poor answer, as indicated in the scoring instructions for the test,³ was scored simply as one poor answer. Each "failure" (i.e. "Nothing At All" checked or no mark made within an A, B, or C section of the blank) was scored also as one poor answer.
- (b) Each poor answer was counted in the same manner as above, but each "failure" ("Nothing At All" checked, or no check) was counted as two poor answers.

In both methods the total number of answers each individual had checked was computed and the percentage of her poor answers was considered to be her final score.

² S. R. Hathaway and J. C. McKinley. Booklet for the *Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press. 1943.

³ Harrower-Erickson and Steiner. *Large Scale Rorschach Techniques*. Springfield, Illinois: Charles C. Thomas. 1944.

The cutting point for both methods of scoring was established at 40%. According to the author of the test, individuals earning a score of less than 40% were to be considered "normal"; those earning scores between 40-59% were to be considered doubtful or questionable risks (the higher the percentage, the less stable the individual); those individuals with scores between 60% and 100% were to be considered highly doubtful.

Discussion of Findings

The original plan for final screening called for all women arriving at the staging area to be given both MCR and MMPI. However, after testing the first group of 181 women, it was decided to omit the MCR as it was the opinion of the staff that the test results appeared to be too ambiguous to be used for predicting future behavior with any degree of reliability.

The results of the two methods of scoring are shown in Table 1.

Table 1

Distribution of Scores Made in the Two Methods of Scoring (Method I, Each Failure Counted as One Poor Answer; Method II, Each Failure Counted as Two Poor Answers)
Multiple Choice Rorschach

% Poor Answers	Method I	Method II
100-104%	0	1
95- 99	1	1
90- 94	0	0
85- 89	2	3
80- 84	2	1
75- 79	0	1
70- 74	1	2
65- 69	0	2
60- 64	3	2
55- 59	1	1
50- 54	7	7
45- 49	3	4
40- 44	6	6
35- 39	6	8
30- 34	17	23
25- 29	20	16
20- 24	18	21
15- 19	44	34
10- 14	24	24
5- 9	16	17
0- 4	10	7
	<i>N</i> = 181	181

Table 2

The Median, Semi-interquartile Range, Mean and Standard Deviation for the Two Methods of Scoring

	N	Md	Q	M	SD
Method I	181	19.60	8.54	24.20	17.35
Method II	181	22.02	9.61	26.3 [^]	19.20

Table 3

Comparison of Scores Made Using the Two Methods of Scoring (Method I, Each Failure Counted as One Poor Answer; Method II, Each Failure Counted as Two Poor Answers)

	Method I	Method II
Total no. above cutting point*	26 (14%)	31 (17%)
No. scores within range 60-104 (highly doubtful)	9 (5%)	14 (8%)
No. scores 40-59 (doubtful)	17 (9%)	18 (10%)
No. scores 0-39	155 (86%)	150 (83%)

* Cutting point = 40.

Table 4

Number of Scores Raised When the Failures Were Counted as Two Poor Answers

	No.	Per Cent
Number of scores raised	65*	36
Number of scores not raised	116	64
Number raised <i>within</i> any given group	55	30
Number raised from one range to <i>another</i>	9	5
Number of scores raised within:		
0-39	46	25
40-59	6	3
60-100	3	2
Number of scores raised from range:		
0-39 to 40-59	5	3
40-59 to 60-100*	4	2
Number of scores raised two ranges	0	

* 1 score was raised to above 100%.

In the cases where the MCR scores were raised from below to above the cutting point as shown in Table 5, only one woman made significant deviate scores on the MMPI. Both her hypomania and schizophrenia scales have critical scores. However, her F or validity score is not only a critical one but sufficiently high to consider the possibility of the individual's having purposely attempted to make a bad set of scores,

particularly since the scores made on the depression and hysteria scales are the only ones not closely approaching the cutting score.

As may be seen from Table 1, when each failure was counted as 1 poor answer, there were 26 women with scores above the cutting point, 9 of whom were "highly questionable," that is, having scores between 60% and 100%. The second method of scoring produced 31 women with scores above the cutting point, 13 of whom were "highly questionable." Considering only those scores at the very extreme in the 1st method of scoring, those between 60% and 100%, even in a random sample, which this is not, it would be surprising to find that 5%, or 9 individuals out of a sample of 181 earned such extremely significant scores unless they were deliberately trying to do so. But in a highly selected group of women, where the motivation and incentive to do well on the test reduces to a

Table 5

Multiple Choice Rorschach Scores Raised Above the Cutting Point by the Second Method of Scoring and the Scores Made by the Same Five Individuals on the Minnesota Multiphasic Personality Inventory

Multiple Choice Rorschach* % Poor Answers Method I Method II		Minnesota Multiphasic Personality Inventory									
		L	F	Hs	D	Hy	Pd	Pa	Pt	Sc	Ma
37%	47%	60	50	48	51	49	56	50	51	52	48
34	43	56	50	39	47	56	51	50	41	39	57
30	47	66	50	37	40	47	42	44	34	39	48
30	40	53	50	41	49	54	40	41	34	40	52
28	44	50	80	62	56	52	68	67	68	75	72

* Cutting point = 40.

minimum the chances of their deliberately trying to earn poor scores, these extreme scores become even more surprising and somewhat less understandable.

By similar reasoning we would ordinarily expect individuals with such highly significant scores as 80% to 100% to produce at least significant deviate scores on other tests for detecting personality disturbances. However, as will be seen in Table 6, those individuals with significant MCR scores did not have *any* significant scores on the 8 major scales of the MMPI; nor did the author or other members of the staff observe any behavior symptomatic of poor adjustment or instability. It might mean, of course, that the MCR is a supersensitive measuring instrument, capable of detecting potential behavior deviations beyond other tests and beyond the elements of selectivity used in this group, except that if this were the case and we started our reasoning with the ten individuals who

made significant scores on the MMPI, then we should expect to have the "more sensitive instrument" showing deviations for these people also. However, inspection of Table 7 does not substantiate such an expectation. Not once, by the first method of scoring, did the "more sensitive instrument" uncover potential maladjustment in the individuals so indicated by the MMPI. On the other hand, by the second method of scoring the only individual who earned a significant deviate major scale score is the woman with the F score of 80 which probably indicates invalidity.

In the ten cases of deviation on the MMPI, as shown in Table 7, it is

Table 6

Significant Multiple Choice Rorschach Scores and the Scores the Same Individuals Made on the Minnesota Multiphasic Personality Inventory

Multiple Choice Rorschach*		Minnesota Multiphasic Personality Inventory†									
Method I	Method II	L	F	Hs	D	Hy	Pd	Pa	Pt	Sc	Ma
95%	95%	70	50	43	47	56	51	50	34	37	48
88	88	63	50	39	46	56	47	49	37	42	66
87	103	53	53	41	36	59	61	50	37	43	57
81	89	63	50	43	53	61	43	50	39	40	50
80	80	50	55	39	36	31	47	50	50	52	68
74	74	63	50	46	38	57	49	53	39	42	54
63	66	50	50	41	42	43	37	38	39	43	57
63	77	53	55	39	63	45	47	56	50	44	39
62	62	50	50	37	44	40	40	50	43	44	52
58	87	50	50	37	40	43	49	44	36	44	59
53	53	56	50	39	42	57	51	56	34	42	54
53	60	60	50	37	38	49	58	47	41	40	66
51	53	56	50	52	51	49	49	62	56	60	68
50	57	56	50	39	44	54	56	41	36	39	48
50	50	56	55	39	55	50	58	56	37	42	54
50	50	66	56	45	49	57	61	59	45	50	52
50	50	56	50	39	49	50	43	65	36	39	43
49	51	50	50	39	38	50	40	41	37	39	63
46	53	70	64	39	49	45	42	35	41	40	54
46	46	63	50	39	40	50	40	47	36	40	57
44	70	56	50	39	47	50	47	53	34	40	50
43	48	80	50	41	47	50	43	56	36	42	43
42	42	53	50	41	36	40	30	41	41	42	50
41	66	60	58	37	47	42	51	44	42	49	48
40	40	63	50	37	55	43	49	47	38	49	50
40	43	53	55	37	55	56	42	50	47	49	37

* All scores on Multiple Choice Rorschach of 40 and above considered significant.

† All scores on Minnesota Multiphasic Personality Inventory of 70 and above considered significant.

interesting to note that eight out of the ten deviations are on the Ma (hypomania) scale. An Ma score of 70 to 75 is not ordinarily interpreted as being necessarily undesirable or maladjustive, providing the rest of the profile is good. And considering the unusual set of motivation factors in this situation, it is not at all surprising, or even "deviational," to find significant scores made on the scale which indicates persons who are over productive in thought and action, who are full of vigor, ambition, plans and enthusiasm. Before volunteering they were all told that life overseas would be more rugged, harder, less convenient, less comfortable than in the States; that there was a great deal of work to be done and that probably they would all be required to put in many extra hours work. In other words, it was put up to them as a challenge; and therefore it probably would be surprising only if there were *no* Ma scores of 70 or over.

Table 7

Significant Minnesota Multiphasic Personality Inventory Scores and the Scores the Same Individuals Made on the Multiple Choice Rorschach Test

L	Minnesota Multiphasic Personality Inventory*									Multiple Choice Rorschach	
	F	Hs	D	Hy	Pd	Pa	Pt	Sc	Ma	Method I	Method II
56	56	45	32	47	44	50	43	38	72	33%	33%
50	50	41	40	49	56	44	43	49	77	30	30
50	80	62	56	52	68	67	68	75	72	28	44
50	53	40	38	52	56	73	38	54	63	27	30
50	50	43	38	56	47	57	52	47	72	24	24
50	50	45	44	49	61	41	54	53	84	19	19
53	50	39	46	47	37	53	49	50	75	17	17
50	53	41	46	56	63	44	50	55	70	17	17
50	50	39	32	36	51	47	49	45	75	16	16
53	50	39	42	42	42	35	38	44	72	10	10

* All standard scores of 70 and above are considered significant.

Conclusion

Since there was no correspondence between the scores made on the Multiple Choice Rorschach (MCR) and the Minnesota Multiphasic Personality Inventory (MMPI), nor any observed behavior which warranted a diagnosis of maladjustment such as the extreme scores made on this test would indicate, it must be concluded that the MCR differentiates something other than it purports to do and that further research and standardization are necessary before the test can be used on a similarly selected sample for the screening of maladjusted individuals.

Received October 8, 1945.

The Relationship Between Subjective Estimates of Personal Adjustment and Ratings on the Bell Adjustment Inventory

Jacob Tuckman

Jewish Vocational Service, Montreal, P. Q.

The recognition that personal maladjustment is a contributing factor to job maladjustment has pointed up the need for the vocational counselor to gather information regarding the ability of the counselee to adjust to others, and the extent to which he may be burdened with personal problems. Such information is sometimes obtainable from the school, parents, social agencies which may have had contact with the counselee or his family, as well as from observation during counseling interviews. In addition, however, it has been found advisable to include in the test battery, consisting of tests of intelligence, achievement, interest, and specific aptitudes, some measure of adjustment. Although many tests under the broad heading of personality, adjustment, and character have been developed, validity has been generally too low to warrant their use for individual diagnosis. These tests are too often dependent upon the honesty and insight of the counselee. Nevertheless, they have been found useful in providing clues to an individual's adjustment, and in serving as points of reference in the interview.

One widely used test to determine an individual's adjustment is the Bell Adjustment Inventory, designed for use with students and adults. The student form, consisting of 140 questions to be answered by "Yes," "No," or "?," gives scores in four adjustment areas,—Home, Health, Social, and Emotional, as well as the total adjustment based on the sum of the four separate scores. The adult form, consisting of 160 questions, yields an "Occupational" adjustment score in addition to the other four measures, but is not scored for this adjustment area if the individual is not or has not been employed. Either inventory can ordinarily be completed in twenty to thirty minutes.

At one point in the testing program at the Cleveland Jewish Vocational Service, when time limitations presented a serious problem, we became interested in determining whether a shorter, more direct approach to the measurement of a counselee's adjustment derived from the Bell Adjustment Inventory could be devised which would be a satisfactory substitute for that test. A set of five statements covering the same areas

and degrees of adjustment as are included in the Bell Adjustment Inventory was drawn up, on which the counselee was to rate his adjustment on a five-point scale. The shortened scale, hereinafter referred to as the Adjustment Questionnaire, is as follows:

ADJUSTMENT QUESTIONNAIRE

Directions:

Under the following headings, underline the answer that you feel applies to you. Compare yourself with other people you know in choosing your answer. There are *No* right or wrong answers.

1. MY HOME LIFE IS:
EXCELLENT GOOD AVERAGE UNSATISFACTORY VERY UNSATISFACTORY
2. MY HEALTH IS:
EXCELLENT GOOD AVERAGE UNSATISFACTORY VERY UNSATISFACTORY
3. MY ABILITY TO GET ALONG WITH OTHER PEOPLE IS:
EXCELLENT GOOD AVERAGE UNSATISFACTORY VERY UNSATISFACTORY
4. MY HAPPINESS IN LIFE IS:
EXCELLENT GOOD AVERAGE UNSATISFACTORY VERY UNSATISFACTORY
5. MY SOCIAL CONTACTS ARE:
EXCELLENT GOOD AVERAGE UNSATISFACTORY VERY UNSATISFACTORY

The statements on Home and Health adjustment presented no difficulties, but it was felt that those on Social and Emotional adjustment needed different phrasing. Statements three and five covering Social adjustment seemed more appropriate than the use of either statement alone. Statement four seemed to cover the Emotional area adequately.

The subjects for the study were 191 high school boys, 200 high school girls, 51 men, and 45 women, who were referred for testing by the counseling and placement departments of the Cleveland Jewish Vocational Service. The school group was enrolled in a college preparatory course in various high schools, or was enrolled in junior high schools normally leading to such a course of study. The group was about equally distributed in grades nine to twelve. Both boys and girls were superior in intelligence as measured by the American Council on Education Psychological Examination for High School Students (1941 and 1942 editions). The adult group consisted of unemployed men and women who were in need of job counseling. The majority were high school graduates. The men were superior in intelligence, the women average, as measured by the Pressey Senior Classification Test. The Adjustment Questionnaire was given before the Bell Adjustment Inventory to half of the subjects and after the Bell to the other half.

For each of the four groups, contingency coefficients for the ratings on the Bell Adjustment Inventory and the Adjustment Questionnaire were computed and corrected for broad groupings. These are presented in Table 1. The corrected contingency coefficients, with the exception of the Emotional adjustment for boys, are surprisingly high. There is little sex difference. The contingency coefficients for the adult groups are higher than the school groups in all areas.

Table 1

Contingency Coefficients for Ratings on the Bell Adjustment Inventory and the Adjustment Questionnaire for Boys, Girls, Men and Women

Adjustment Area	Boys N = 191		Girls N = 200		Men N = 51		Women N = 45	
	C	Corr. C*	C	Corr. C*	C	Corr. C*	C	Corr. C*
Home	.529	.59	.550	.62	.657	.74	.679	.76
Health	.412	.46	.366	.41	.722	.81	.587	.66
Social (3)	.536	.60	.423	.48	.580	.65	.664	.75
Social (5)	.449	.50	.549	.62	.537	.60	.675	.76
Emotional	.281	.32	.526	.59	.599	.67	.534	.60

* Corrected for broad groupings by dividing C by .89. See Peters, Charles C., and Van Voorhis, Walter R. *Statistical procedures and their mathematical bases*. Pp. 393-399.

Table 2 gives the per cent of each of the four groups whose ratings on the Bell Adjustment Inventory and the Adjustment Questionnaire are identical. These vary from 22% for the Health adjustment for boys, to 44.4% for the Home adjustment for women. The groupings regarding the degree of adjustment for the Bell Adjustment Inventory are fairly arbitrary, since one point in score may change an individual's rating from one category to another. For practical purposes, we are primarily interested in knowing whether a person's adjustment is average, above average, or below average. Table 3 presents the per cent of the four groups who were above average, average, and below average, for the various adjustment areas on both measures. As may be expected, the per cent of the groups falling within the same category on both measures is considerably higher than that obtained on a five-point scale. These vary from 37.2% for the Health adjustment for boys to 64.4% for the Social adjustment (5) for women. In all areas, the per cents are higher for girls and women than for boys and men but these differences are not statistically reliable.

Table 2

Per Cent of the Four Groups Whose Adjustment Rating on the Bell Adjustment Inventory and the Adjustment Questionnaire was Identical

Adjustment Area	Boys %	Girls %	Men %	Women %
Home	27.2	37.5	35.3	44.4
Health	22.0	29.0	35.3	33.3
Social (3)	39.8	40.5	27.5	24.4
Social (5)	42.9	43.0	37.3	40.0
Emotional	28.3	33.5	33.3	40.0

Table 3

Per Cent of the Four Groups Whose Adjustment Rating was Above Average, Average, and Below Average on the Bell Adjustment Inventory and the Adjustment Questionnaire

Adjustment Area	Boys %	Girls %	Men %	Women %
Home	45.0	57.0	56.9	57.8
Health	37.2	40.0	49.0	55.6
Social (3)	55.5	57.5	37.3	60.0
Social (5)	58.1	58.5	51.0	64.4
Emotional	36.6	45.0	49.0	51.1

In comparing the adjustment on both measures, each of the four groups has rated itself more favorably on the Bell Adjustment Inventory. Of special interest therefore, are cases where the adjustment on the Adjustment Questionnaire is poorer than that given by the Bell Adjustment Inventory. In some cases there is a wide discrepancy,—e.g., *excellent* on the Bell Adjustment Inventory, and *very unsatisfactory* on the Adjustment Questionnaire. The per cent of each group whose adjustment is poorer on the latter is presented in Table 4, and shows variations from 4.7% for the Health adjustment for boys to 28.9% for Social adjustment (5) for women. With the exception of the Health adjustment for women, girls and women tend to have a poorer adjustment on the Adjustment Questionnaire than do boys and men, but these differences are not significant.

Table 4

Per Cent of the Four Groups Whose Adjustment was Poorer on the Adjustment Questionnaire than on the Bell Adjustment Inventory

Adjustment Area	Boys %	Girls %	Men %	Women %
Home	8.4	11.0	7.8	11.1
Health	4.7	6.5	11.8	11.1
Social (3)	11.0	19.5	5.9	15.6
Social (5)	16.2	22.5	19.6	28.9
Emotional	11.5	13.5	19.6	22.2

Summary

The data indicate that it is possible to obtain a fairly good estimate of an individual's adjustment by merely using a few simple direct questions. For practical purposes, the correlation between the two measures is high

enough to warrant the substitution of the Adjustment Questionnaire for the Bell Adjustment Inventory if there is not sufficient time to give the latter.

The incidence of a sizeable proportion of cases whose adjustment on the Adjustment Questionnaire is poorer than on the Bell Adjustment Inventory is especially noteworthy. This does not mean that the Bell Adjustment Inventory is invalid, but that an individual's subjective judgment as to what he estimates his own adjustment to be is also important psychologically. Certainly, it is desirable, if time permits, to include a questionnaire of this sort along with the Bell Adjustment Inventory to give the counselor additional insight into the situation.

Received November 5, 1945.

Item Difficulty of Some Wechsler-Bellevue Subtests

A. I. Rabin, J. C. Davis, and M. H. Sanderson

New Hampshire State Hospital, Concord, N. H.

It is the experience of every psychometric examiner with most testing scales that some items which are supposed to be "easy" in view of their placement in the beginning of the scale are failed, while the more "difficult" ones, placed toward the end of the scale, are passed. The occasional occurrence of such inconsistencies does not disturb the examiner and is attributed to the idiosyncracies of the testee and to the peculiarities of his mental development, whether in the normal or abnormal range. This is also to be expected in view of the fact that final placement of items in order of difficulty is usually based on numerical summaries and statistical data of large standardization groups which mask individual patterns. However, when such "inconsistencies" occur with some degree of regularity and consistency, the correctness of the order of items is to be questioned, at least, for the sample of the population involved.

Extensive experience in the application of the Bellevue Intelligence Scales with normal and abnormal adults raises the question of the correct placement of items within the several subtests. We have observed, especially in such subtests as information, picture completion, comprehension, similarities and others, that some of the items that appear in the first part of the respective subtests are much more difficult for our subjects and are failed with greater frequency than those appearing in the middle or last part of those tests. Quantitative proof and substantiation of these empirical notions appeared desirable.

Wechsler's last edition of the *Measurement of adult intelligence*¹ shows an implicit recognition of the need for analysis of item difficulty. In fact, Wechsler adopts modifications in the order of presentation of the Information questions, based on unpublished data communicated by Altus (p. 172).

An *intra-test* analysis of item difficulty would accomplish two major services in the clinical situation. In the first place, a more correct and statistically justified arrangement of items would make the test a more efficient tool. Several consecutive failures on successive items would actually mean little chance for success beyond that level. Thus time

¹ Wechsler, D. *The measurement of adult intelligence* (Third edition). Baltimore: The Williams and Wilkins Co., 1944.

would be saved and further useless questioning obviated. Secondly, it might be less discouraging to some sensitive testees who have to experience failures before reaching the actual limit of their capacities.

The Subjects

The present study is based on three hundred Wechsler-Bellevue records of individual examinations of normal persons, drawn from three sources: 1. 100 subjects were student nurses at the New Hampshire State

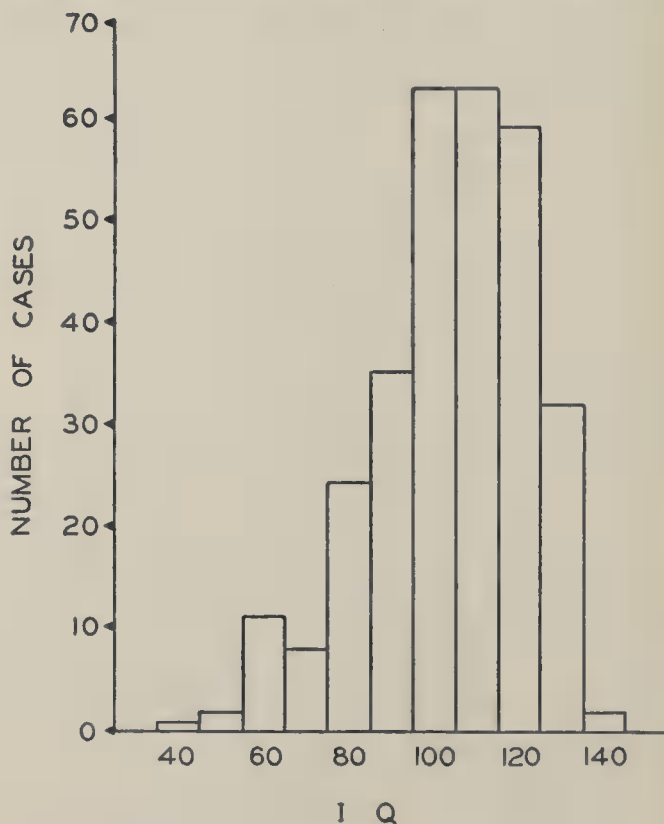


FIG. 1. Numerical distribution of IQ's of entire group.

Hospital School of Nursing; 2. 40 were members of a Conscientious Objectors unit stationed at the Hospital; and 3. 160 were vocational guidance cases. Some were referred to the Psychology Department for vocational advice, while others were private clients of the senior author. The mean IQ (full scale) for the entire group is 104.4.

It is obvious, from Figure 1, that our group is better than average intellectually. Very few extremely retarded cases were included (5% feeble-minded and about the same percentage of borderline mentality). The overwhelming majority consisted of individuals with better than dull

normal mentality. Though the distribution covers a wide range, it is clearly skewed to the left.

Procedure

An analysis of the difficulty of the items in six subtests of the Wechsler-Bellevue Scales was undertaken. The items of the information test (25 items) as mentioned earlier, on the basis of clinical observation and experience, are certainly not arranged in order of difficulty. Similarly, two other verbal subtests exhibited considerable scatter, i.e., comprehension (10 items) and similarities (12 items). Of the remaining regular verbal subtests, the arithmetic test shows little irregularity, while the digits (forward and backwards) test does not lend itself to a difficulty analysis, since the gradation of difficulty depends on quantitative (number of digits repeated) rather than differences in difficulty of individual items.

Three performance subtests lend themselves readily to analysis: picture completion (15 items), picture arrangement (6 items) and block designs (7 items). The object assembly test was not included, since only three items comprise the test. These seem to be arranged in order of difficulty. Because of the quantitative, uniform nature of the test, digit symbols were not included.

Thus, the following analysis is applied to three Verbal and three Performance tests which can be submitted easily to item by item scrutiny. The procedure was to count the number of subjects (of the total group of 300) passing each item on each test. The corresponding ranks of items were then computed on the basis of the ease with which they were passed by our subjects. Thus rank 1 is assigned to an item within a test which showed the highest number of successfully passing subjects; rank 2, next to the highest number, etc.

The Data

Wechsler's original order of the items of the Information test is given in the first column of Table 1. The fourth column presents the revision of the order of the first 20 items based on Altus's data and published in the third edition of Wechsler's book.² Our rank order for the items, based on the data in the second column, may be found in column 3 of Table 1. The ranks in the parentheses are based on the actual data and were caused by the special circumstances referred to in the footnote to the table. Otherwise, the rank order is as indicated. The first six items agree perfectly with the order suggested in Wechsler's recent revision.

² *Op. cit.*

Table 1
Data on Items of the Information Test

Question No. (Wechsler)	No. of Passes	Rank Order	Revised Order (Altus)
1	227	(9)* 1	1
2	271	(3) 4	4
3	254	(4) 5	5
4	272	(2) 3	3
5	291	(1) 2	2
6	237	(5) 6	6
7	211	13	7
8	228	(8) 9	16
9	229	(7) 8	9
10	226	10	10
11	143	18	12
12	231	(6) 7	11
13	214	12	14
14	78	20	17
15	168	16	20
16	181	14	8
17	218	11	13
18	118	19	15
19	175	15	19
20	150	17	18
21	45	23	21
22	49	22	22
23	66	21	23
24	8	25	24
25	19	24	25

* Most of the subjects were tested during the late President Roosevelt's 2nd and 3rd terms; hence, the difficulty in naming his predecessor which is required on this item.

The remainder show considerable variation. It is quite obvious that item 14 (discoverer of the North Pole) is badly misplaced since it ranks 20th in difficulty. The same is true of items 11 and 17. Lesser discrepancies can be found in the remainder of the order.

Roughly speaking, the test may be subdivided into three sections: questions 1-10, 12 and 17 which are easiest and are passed by more than $\frac{2}{3}$ rds of our population. Then questions 14 and 21-25 are by far the most difficult and are passed by approximately $\frac{1}{3}$ th of our population. The remaining items, passed by one-half of our subjects, do not provide a very gradual transition to the most difficult items.

It is quite clear that unless the revised order of items is followed in the examination, the rule of discontinuing the questions after five successive failures may cause too many errors.

A difficulty analysis of the other two verbal subtests is given in Table 2. Here, too, some revision of order is desirable. Items 2 and 8 of the comprehension test, which are practically of the same difficulty, are 5 ranks apart. Items 9 and 10 also need to change places. The sudden drop in the last two items is quite interesting. While nearly 85% of the subjects are able to pass the first 8 items, only 45% and 65% of cases are able to pass items 9 and 10, respectively. Here, too, a more gradual drop would be desirable. When using the scale, however, if any of items 6, 7, or 8 are failed, no successes on the last two items can be expected and therefore the time required for questioning may be saved.

Table 2
Item Analysis of Two Verbal Tests

Comprehension											
Question	1	2	3	4	5	6	7	8	9	10	
No. of Passes	286	253	286	265	280	245	237	254	134	194	
Rank Order	1.5	6	1.5	4	3	7	8	5	10	9	
Similarities											
Items	1	2	3	4	5	6	7	8	9	10	11 12
No. of Passes	282	289	289	284	291	173	151	216	181	160	63 72
Rank Order	5	3.5	3.5	1	2	8	10	6	7	9	12 11

Minor changes in the similarities test are also suggested by the analysis of item difficulty. There does not seem to be a great deal of variation in the difficulty of the first five items. The following five items with the exception of No. 8 again show a considerably higher level of difficulty. The last two items show a level of difficulty several times higher than the group of items preceding it. In summary, the test may be subdivided into three blocks with a rather sudden transition from one to another: Block One, consisting of items 1 to 5, with about 95% of the subjects passing; Block Two, consisting of items 6, 7, 9 and 10, with about 55% of the subjects passing; and Block Three, consisting of the remaining items (11 and 12), with only about 23% successes. The only seriously misplaced item is No. 8, which should be placed 6th in order.

The results of an analysis of the three performance tests are given in Table 3. The items in the picture completion test show varying degrees of misplacement all along the line. If rearranged in the obtained rank order based on the number of subjects who passed each item, there would be a fairly gradual increase in difficulty. Of course, ordinarily all items are administered. However, time and effort may be saved if the test is

Table 3
Analysis of Three Performance Tests

Picture Completion															
Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of															
Passes	291	278	271	233	173	277	185	242	275	198	105	237	126	76	123
Rank															
Order	1	2	5	8	11	3	10	6	4	9	14	7	12	15	12

Picture Arrangement							
	Picture						
Sets	1	2	3	4	5	6	
No. of							
Passes	295	265	271	207	188	184	
Rank							
Order	1	3	2	4	5	6	

Block Designs								
	Designs	1	2	3	4	5	6	7
No. of								
Passes	293	284	287	245	262	167	146	
Suggested								
Order	1	3	2	5	4	6	7	

discontinued after 3 or 4 consecutive failures, providing the items are presented in order of difficulty.

Little criticism may be levelled against the picture arrangement test. There are minor but insignificant variations. No changes in order can be considered essential.

The third part of Table 3 presents the results for the Block Design tests. According to these data, an interchange of places between designs 4 and 5 would be quite desirable. Otherwise, the order of their arrangement appears to be quite reasonable and characterized by increasing difficulty.

Substantially similar results and, consequently, substantially similar suggestions for intra-test rearrangements were obtained for 1,000 State Hospital patients. Our suggestions, however, are designedly not based on the findings with this much larger sample of a wider age range and wider range of levels of ability, since such an analysis of item difficulty may reflect the "scatter" and selective performance of the patients and may not hold strictly for a "normal" population. The present results are nevertheless corroborated by those findings and are given greater strength and conclusiveness by them.

Summary

Clinical experience during the administration of the Wechsler-Bellevue scales prompted a detailed quantitative analysis of the difficulty of items based on 300 records of normal individuals. The analysis was confined to the following six subtests which lent themselves to such statistical treatment: information, comprehension, similarities, picture completion, picture arrangement, and block designs. The suggested changes in the original order of item presentation are summarized in Table 4. The results are also largely substantiated by findings on 1,000 psychiatric patients not included in this study. It is felt that the findings may help speed up the administration of the test and avoid a frequent excessive sense of failure on the part of less gifted examinees.

Received October 8, 1945.

The Relationship Between Knowledge of Human Development and Ability to Use Such Knowledge *

John E. Horrocks

Ohio State University

A practitioner is one who applies knowledge and skill to practical situations. It is commonly assumed that ability to apply stems from a background of knowledge and experience. Hence, to apply is to know,—to be steeped in principles and facts germane to the discipline being applied. That more than knowledge of facts and principles enters the picture is admitted. A pressing question in clinical practice and in teaching is the relationship that knowledge of facts and principles bears to ability to apply those facts and principles. This question has remained largely unanswered.

Practically it would seem to follow that a person may not apply what he does not know. For that reason, and for the added reasons of administrative expediency, time, and expense, courses for the training of clinical psychologists and teachers have tended to concentrate on factual retention. Thus, those who in their professional capacities are called upon daily to deal with the intricacies of human behavior, bring to their work a background of factually oriented courses in psychology and education.

Purpose of the Study

It is the purpose of this study to examine the relationship existing between knowledge of facts and principles and ability to apply such facts and principles to one area of human development,—adolescence. A psychologist or teacher has to deal with human behavior in complex life situations. He must diagnose and take remedial action in circumstances complicated by inter-related and constantly evolving situations. He must apply what he knows about human behavior to the occasion which confronts him, bearing in mind all the while the antecedents of the occasion and the consequences of his decisions. This study will analyze the performance of a selected population, whose knowledge of the facts and principles of adolescent development has been tested, when confronted with a complex situation in which they are given an opportunity to make a diagnosis and select proper remedial procedure.

* Grateful acknowledgment is made to Dr. Maurice E. Troyer of Syracuse University.

Reliability and Validity of Instruments Used

In carrying out the purpose of the study, four tests were constructed. The first was a criterion test which measured knowledge of fact and principle about adolescent development. Second, three case study tests were constructed which measured ability to apply the facts and principles of adolescent development. Each case study was divided into three parts. After each part the student was given an opportunity to reveal his ability to diagnose difficulties and select appropriate remedial procedures. Answers to the case study tests were scored with a key based on a weighted composite of expert opinion.

The criterion test was designed to cover the major principles and facts ordinarily included in texts and adolescent psychology courses. The subject matter coverage of the three case studies paralleled that of the criterion test. Each case centered around a different problem in adolescence,—social, academic, and emotional.

The reliability coefficient obtained by the split-half method for the criterion test was $.91 \pm .017$. The validity of the criterion test was based on its reliability, internal consistency, coverage, construction, and keying.

The split-half correlation obtained for the *Case of Barry Black* was $.79 \pm .038$; for the *Case of Sam Smith*, $.73 \pm .046$; and for the *Case of Connie Casey*, $.77 \pm .041$. The validity of the three case studies rested upon construction and coverage, expert scoring, item consistency, reliability, and utility.

Procedure

The criterion test and the three case studies were administered to populations of college juniors, seniors, and graduate students taking courses having to do with adolescent behavior, educational psychology, and mental hygiene.

The *Case of Barry Black* was administered to a group of 100 college students, composed of 90 liberal arts and teachers college juniors and seniors, 7 graduate students, and 3 nurses in training. Thirteen of the 100 were experienced teachers. All people selected were midway through a college course having to do with human behavior.

During the class meeting prior to the administration of the *Case of Barry Black*, the class answered the criterion test.

The same procedure was followed in the administration of the cases of *Sam Smith* and *Connie Casey* to groups of 100 each. For the *Case of Sam Smith* the group selected consisted of 42 teachers college juniors and seniors, and 58 liberal arts college juniors and seniors. For the *Case of Connie Casey*, 9 graduate students, 15 registered nurses (taking under-

graduate work in public health nursing), and 76 liberal arts and teachers college juniors and seniors were selected.

Product-moment coefficients of correlation were computed between the criterion test and the whole, diagnosis, and remedial scores for each of the three case studies. The correlations existing between the criterion test and the three case studies are given in Table 1.

Table 1

Coefficients of Correlation Existing Between the Criterion Test and the Three Sections of *Black*, *Smith*, and *Casey*

Variable	Whole	Diagnosis	Remedial
Black-criterion	.46 \pm .078	.49 \pm .075	.29 \pm .091
Smith-criterion	.41 \pm .083	.40 \pm .084	.30 \pm .091
Casey-criterion	.26 \pm .093	.24 \pm .094	.26 \pm .093

The three populations used above may be considered by virtue of selection and composition to be representative of a particular group of college students. They were all parts of larger populations,¹ and they consist, for the most part, of college juniors and seniors with approximately the same general educational experience and age. In all cases the tests were given midway through the course. For that reason it may be assumed that differences existing among correlations between the criterion test and the three case studies are due to chance or differences in the case studies rather than to differences in the populations being considered.

An examination of Table 1 indicates that, *Barry Black* and *Sam Smith* correlate more highly with the criterion test than does *Connie Casey*. It will also be noted that the correlation of *Barry Black* and *Sam Smith* with the criterion test are approximately equal.

Where the diagnosis section is concerned, the same general trend may be noticed. *Connie Casey* correlates less well with the criterion test than does either of the others. *Barry Black* and *Sam Smith*, correlate with the criterion test to about the same extent.

The remedial sections of all three case studies correlate with the criterion test to about the same extent.

Second Administration of *Connie Casey*

The case study tests used for the experiment described above were administered in mimeographed form. As a further check the *Case of*

¹ The populations were selected from existing classes studying adolescent behavior at two state and two private universities. If a given population number, as 100, was required and the available class population was in excess of 100, the extra cases were thrown out from the bottom of the pile of cases being scored.

Connie Casey was revised and printed. *Connie Casey* was chosen for printing because its coverage appeared to be wider than that of either of the other two cases.

The printed case was presented to two groups. The first group, to be known as Group A, consisted of 47 randomly selected graduate students. All were experienced teachers, and all had had previous courses in psychology.

Group B consisted of 69 randomly selected teachers college seniors. None were experienced teachers, but all had completed their practice teaching and were finishing their professional training. All had completed a course in adolescent development during the previous school year.

Coefficients of correlation were computed between the criterion test and each of the three sections of the revised *Case of Connie Casey* for Groups A and B. Table 2 shows the correlations between the criterion test and the case study for Groups A and B.

Table 2
Correlation Between the Printed Edition of *Connie Casey* and the Criterion
Test for Groups A and B

Section	Group A (Graduates)	Group B (Undergraduates)
Whole Test	.28 \pm .136	.16 \pm .12
Diagnosis	.35 \pm .129	.27 \pm .11
Remedial	.02 \pm .147	-.04 \pm .12

Here, again, was a positive but slight relationship between the criterion test and the case study. In all cases the standard error for the populations taking the revised test was greater than for that taking the original test because of the greater number in the original group. The following relationships emerge in a comparison of Groups A and B with the original population:

1. The correlation of the whole test with the criterion tests was .02 higher for Group A than for the original population.
2. The correlation of the whole test with the criterion test was .10 lower for Group B than for the original population.
3. The correlation of the diagnosis section with the criterion test was .11 higher for Group A and .03 higher for Group B than for the original population.
4. The correlation of the remedial section with the criterion test was .24 lower for Group A and .30 lower for Group B than for the original population.

The whole and diagnosis increases could well be due to chance, but the remedial correlation drop would appear to be a significant one whose answer might possibly exist in the changed nature of the population or of the test. A drop of .30 is higher than could be explained by chance, even at three sigmas.

In other words, for both Groups A and B the ability to make a diagnosis correlated more highly with the criterion test than did the diagnostic ability of the original population. For both Groups A and B the ability to suggest a remedial plan correlated much less highly with the criterion test than did the remedial ability of the original population.

Results on the remedial section, for Groups A and B, lead to the conclusion that the printed form of *Connie Casey* has certain slightly different potentialities than has the original mimeographed form.

From the point of view of the present study, however, the differences found on the revised edition of *Connie Casey* are not significant. In no case has any correlation been as high as .50, and the trend of the results from the revised *Connie Casey* indicates less of a positive relationship than would have been inferred from the original correlations obtained between the criterion test and the three original case studies. The conclusion is inescapable that the case studies (whole, diagnosis, and remedial), while measuring certain aspects in common with the criterion test, are at the same time measuring something which the criterion test fails to measure, and vice versa.

Comparison with Intelligence and Class Marks

Though not directly pertinent to the purposes of this study, it was believed advisable to compute the correlation existing between the case study tests and intelligence, and between the cases and final marks in a course in adolescent psychology. The Ohio State University Psychological Test, Form 22, was administered to a randomly selected group of 61 teachers college juniors, and the group was given the revised edition of *Connie Casey*. The relationship existing between the various sections of the *Case of Connie Casey* and intelligence are given in Table 3.

Table 3

Relationship Between Scores on *Connie Casey* and the Intelligence of a Selected Group of Subjects

OSPE vs. Whole Test.....	.23 ± .122
OSPE vs. Diagnostic Section.....	.10 ± .127
OSPE vs. Remedial Section.....	.25 ± .121

Results shown in Table 3 would appear to indicate a very small relationship between intelligence and the case study tests with this particular group.

It may be concluded from the foregoing correlations that the test is measuring factors other than intelligence, and that a more intelligent person might or might not do as well as a less intelligent one. Intelligence being, of course, in this case, as measured by the OSPE.

At this point it might be asked if it would be possible for a dull normal person to receive as high a rating on the test as a superior person, since intelligence apparently plays a minor part. The answer is "no." It must be remembered that the population involved was made up of successful college students, and therefore of persons of superior or near superior intelligence. The finding is, then, that given a basically good intelligence, added increments of intelligence in the superior range do not add to ability to succeed on the test in question. From the foregoing correlations it might also be tentatively assumed that intelligence may be minimized as a factor in comparisons between the case study and criterion tests.

A coefficient of correlation was secured between the final marks of 59 teachers college juniors in a course in adolescent development and their scores on the *Case of Connie Casey*. The case was not used in determining final marks. The coefficient of correlation was $.38 \pm .112$, which shows a slight positive relationship. The coefficient of correlation found between 61 juniors' marks in the same course and their scores on the criterion test was $.50 \pm .095$, and increase of .12 over the case study's correlation with class marks. The difference is not significantly large, but for the people used in making the comparison there was more agreement between the criterion test and the final mark than between the case study and the final mark. In considering this relationship it must be remembered that the course grades were based on tests somewhat similar to the criterion test. This might help to explain the difference of .12 between the criterion test and the case study.

Inter-Relationships between the Case Studies

The question next arises as to the inter-relationships between the case studies. If it is to be assumed that each one is measuring ability to apply fact *per se*, then it would be expected that a high positive correlation would exist. On the other hand, in constructing the case studies, each case was made to deal with a different aspect of adolescent development, and the question might arise as to whether ability to apply knowledge about an emotional problem would indicate ability to apply knowledge about a social or physical problem. There is also the question as to

whether ability to apply knowledge about a school situation would indicate equal ability to apply knowledge about a home or community situation. If certain basic factors are involved, a positive, though not necessarily high correlation might be expected.

As a matter of fact, with 67 cases the coefficient of correlation between *Barry Black* and *Sam Smith* was $.55 \pm .09$. With 68 cases the correlation between *Barry Black* and *Connie Casey* was $.39 \pm .10$. The correlation between *Sam Smith* and *Connie Casey* was $.62 \pm .09$.

Summary and Conclusions

It has been the purpose of this study to find the relationship existing between knowledge of fact and principle about human development on the one hand, and ability to apply those facts and principles on the other. Four tests were constructed. First was a criterion test which measured knowledge of fact and principle about adolescent development. Second, three case study tests were constructed which measured ability to apply the facts and principles of adolescent development.

The criterion test and the three case studies were administered to 300 college upperclassmen and graduate students taking courses having to do with adolescent behavior. As a check a printed revision of one of the case studies together with the criterion test was administered to a group of 47 graduate students and to a group of 69 undergraduates. Interrelations existing between the criterion test and the case studies were analyzed. Subsidiary analyses were made of the relationship existing between one of the case studies and intelligence, and final grades in a course in adolescent development.

As a result of the study the following conclusions are tentatively made:

1. It would appear that knowledge of facts and principles about adolescent behavior are positively but not highly related to ability to diagnose as measured by the case study tests.
2. It would appear that knowledge of facts and principles about adolescent behavior are positively but not highly related to ability to identify appropriate remedial procedures as measured by the case study tests.
3. Given intelligence enough to pursue college work, added increments of intelligence appear to show a very slight positive relationship to ability to apply facts and principles of adolescent development. The same relationship holds true for ability to recognize the facts and principles of adolescent development after having studied them in a college classroom. The group used in studying these relationships were, however, a select group in that they had gone through a rigorous selection

program before being admitted to teachers college. The median percentile rank for the group (on the basis of the Ohio State University Psychological Test Norms) was 79. No one was under the 50th centile. Hence, the above conclusion should be qualified by noting that the low relationship was found within a comparatively narrow segment of the range of intelligence.

4. The relationship between success in a course in adolescent development as indicated by a final grade and ability to apply facts and principles of adolescent development was positive but small as measured by the instruments used in this study. The ability to recognize facts and principles was not highly related to success in the same course, although it was greater than the ability to apply facts and principles.

5. There does not appear to be an ability *per se* to apply facts and principles of adolescent behavior, but rather ability to apply various facts and principles to varying situations. It may be assumed, however, that there are common factors underlying the various specific abilities. This conclusion is drawn, in part, from the varying performances revealed on the three case studies, each dealing with different aspects of behavior.

6. Insofar as traditional courses in human growth and development have used factual learning as the only method of preparing their students to diagnose and institute remediation in a life situation, they appear to have served their function less adequately than might otherwise be possible. The findings of this study may well challenge the assumption in any course in psychology or teacher education that knowledge of fact and principle necessarily leads to effective or intelligent application of fact and principle.

7. It would appear from the foregoing conclusions that measurement devices traditionally used in teacher education and psychology do not satisfactorily measure ability to apply facts, however well they may measure knowledge of facts themselves. This conclusion is one that might well have been expected when one considers the results of the large number of studies between learning and transfer of training.

Received November 16, 1945.

Keysort Method of Scoring the Minnesota Multiphasic Personality Inventory

Capt. Morse P. Manson, AGD, and Capt. Harry M. Grayson, AGD

Psychologists, MTOUSA Disciplinary Training Center

The Minnesota Multiphasic Personality Inventory (MMPI), "a psychometric instrument designed ultimately to provide, in a single test, scores on all the more important phases of personality,"¹ has been receiving widespread clinical recognition and application. It contains 550 different statements about the behavior, attitudes, and interests of the person being tested, each one appearing on a separate card. The administration is very simple, the examinee being required to sort the cards into three groups *as it applies to him*. However, one practical difficulty encountered in the use of the MMPI is the lengthy method of scoring. ". . . an enlisted man . . . trained in the work . . . usually required twenty to twenty-five minutes for the recording and scoring of a test."² An improved method of scoring which minimizes clerical error, eliminates entirely the recording of individual items, and reduces the scoring time to less than ten minutes has been developed and used in the Personnel Evaluation Department of the MTOUSA³ Disciplinary Training Center.

Rationale

Examination of the MMPI scoring keys reveals that test items may appear on one or more keys. High scores are indicative of mental disturbance. Most of the items (299) are scored only if answered in a "deviate" or infrequent manner. These may be termed *pure deviate* items and appear as X-items on the original scoring keys. Other items (55) are scored only if answered in a "non-deviate" or frequent manner. These may be termed *reverse deviate* items and appear as O-items on the original scoring keys. A final group of 12 items, which are called XO items for identification, are scored on some subtests if answered in a deviate manner and on other subtests if answered in a non-deviate

¹ Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota multiphasic personality inventory*. N. Y.: The Psychological Corporation, 1943.

² Leverenz, Major C. W. Minnesota Multiphasic Inventory, an evaluation of its usefulness in the psychiatric services of a station hospital. *War Medicine*, 1943, 4, 618-629.

³ Mediterranean Theater of Operations, United States Army.

manner. They may be termed *mixed* items and appear either as X- or as O-items on the different original scoring keys. These three types of items require different treatments in scoring.⁴

In the original MMPI scoring method, the CANNOT SAY items are entered in the appropriate cells on the score sheet with a heavy diagonal line and are then discarded. The TRUE and the FALSE groups are conveniently separated into *deviate* and *non-deviate* packets and only the *deviate* answers are used in recording scores. For the cards filed as TRUE items, this group has the lower righthand corners cut. For the cards filed as FALSE items, the lower lefthand corners are cut.

The deviate cards are recorded on the score sheet by an X for each item placed in the appropriate cell. The various subtest scoring keys are then applied, and each item on the score sheet appearing next to an X on the key is counted. Each *blank* item appearing next to a O on the key is likewise counted. The CANNOT SAY items are not credited on any of the keys, but their diagonal line entries are added separately to yield a Question (?) score.

The improvements in scoring the MMPI involve an adaptation of the McBee Keysort System, used by the Army in connection with WD AGO Form No. 20, the Soldier's Qualification Card. This system permits the rapid selection of cards for any desired trait or combination of traits. The four sides of each card are lined with holes, each of which represents a number, trait, or specified item, identified and coded. To indicate the possession of designated characteristics, the proper holes are cut out with a U-shaped notch between the hole and the edge of the card. The insertion of a rod or needle through the proper hole in a stack of cards will release all notched or desired cards.

As applied to the MMPI, the cards are notched in a similar fashion, enabling their rapid removal (and scoring by simple count) for each of the subtests. (See Figure 1.)

The coded holes for the subtests have been arranged counter-clockwise on the reverse side of the cards, following the order of the MMPI

⁴ A large group of items (184) play no part in the scoring on any of the subtests, and are counted only in the CANNOT SAY score. These may be of ultimate significance on new or revised keys. At present, however, they unduly increase test administration and scoring time and may be set aside without invalidating the test to any extent. The twelve XO items, most of which appear on the Hypomania Scale, could also safely be eliminated without harming the test, at the same time greatly simplifying the scoring process. Perhaps a better procedure would be to have the XO cards scored only as X or as O cards depending upon the major role of the particular items. This simplified scoring would more than compensate for the imperceptible change in norms over the various subtests. However, the scoring procedure described in this paper is based on the use of all 550 cards.

score sheet, except that there is no hole for the CANNOT SAY (?) items since the question score is obtained merely by counting the cards. The order of tests on the MMPI score sheet is as follows: (see Figure 2).

- | | |
|-----------------------|-------------------------------------|
| 1. ?—Question score | 7. Pd—Psychopathic deviate |
| 2. L—Lie score | 8. Mf—Interest (masculine-feminine) |
| 3. F—Validity score | 9. Pa—Paranoia |
| 4. Hs—Hypochondriasis | 10. Pt—Psychasthenia |
| 5. D—Depression | 11. Sc—Schizophrenia |
| 6. Hy—Hysteria | 12. Ma—Hypomania |

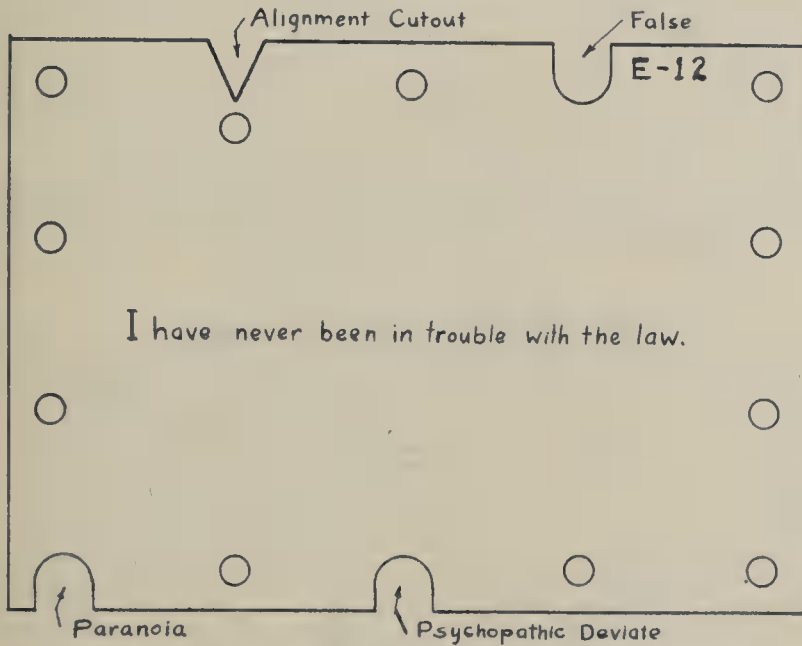


FIG. 1. Typical card, front view.

Method

All MMPI Cards are cut as listed in a Master Coding Chart.⁵ An off-centered triangular cut at the top of each card aids in their proper alignment and prevents possible reversal from the upright, front-face position. The conversion from the original system to the punched-card system of scoring is made as follows:

- a. The X-items or pure deviate items are notched for their respective subtest holes in accordance with the original scoring keys.
- b. The O-items or *reverse deviate* items, since they are scored only if answered in a frequent or non-deviate fashion, have the TRUE and the

⁵ Copies of the Master Coding Chart may be obtained from Morse P. Manson, 8212 Blackburn St., Los Angeles 36, California.

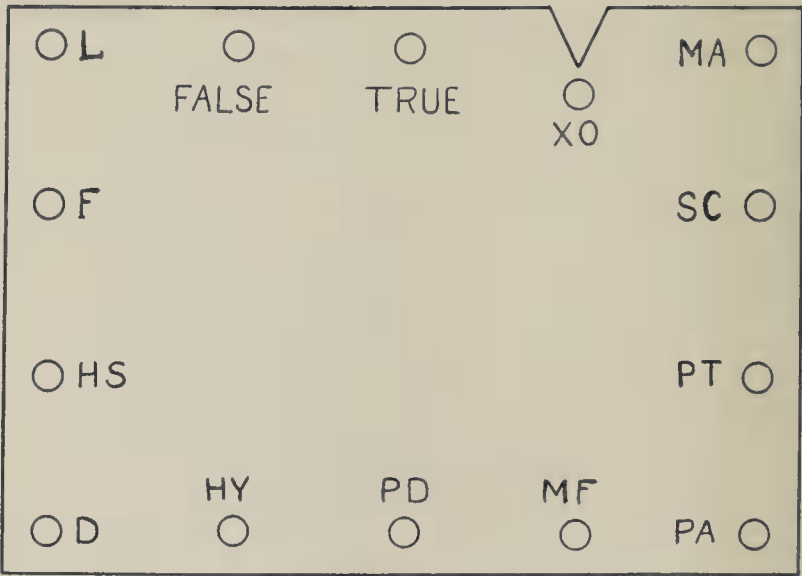


FIG. 2. Guide card (reverse view). (Note: The guide card can be used to aid in accurate needling of the cards. This card is always placed on the reverse side of the stack of cards to be needled.)

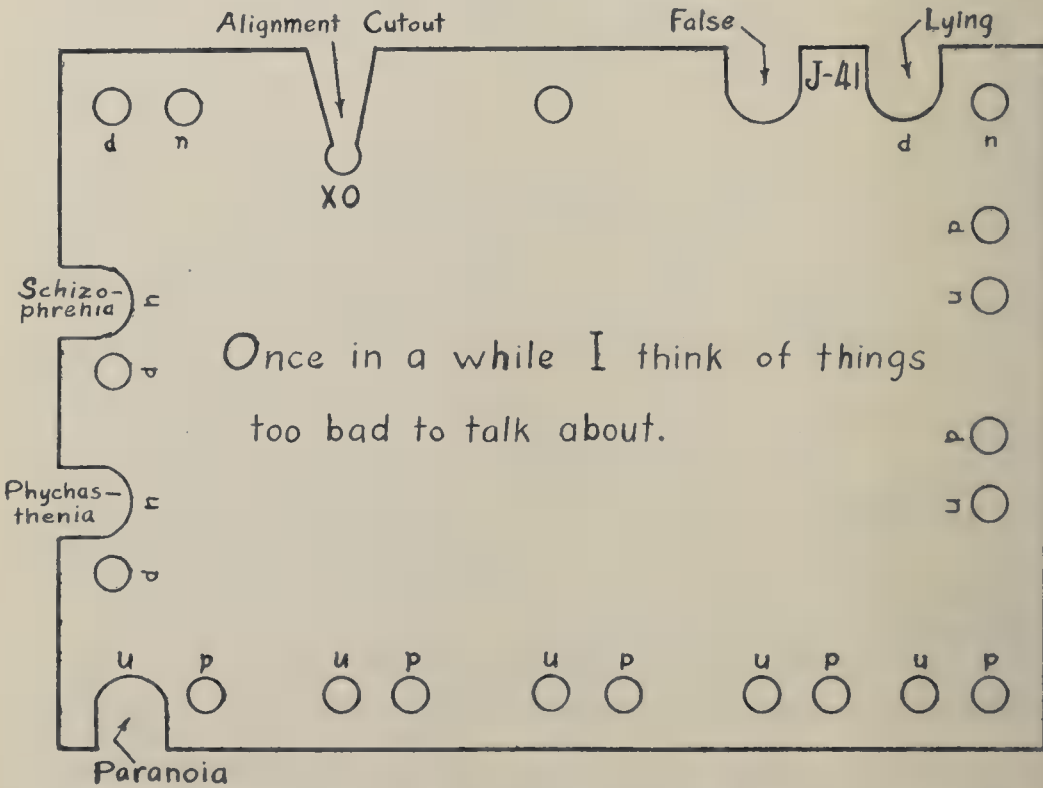


FIG. 3. XO card (front view). (Note: The letters "d" and "n" appear only on reverse side of every XO card.)

FALSE holes notched in the opposite way from the X-items. (These TRUE and FALSE holes are in place of the cut-off corners on the cards.) For example, O-items receiving credit if answered as TRUE, for scoring purposes, are notched as if answered as FALSE. Conversely O-items receiving credit if answered as FALSE, for scoring purposes, are notched as if answered as TRUE. (See Master Coding Chart and Table 1.)

c. The XO items, which on one test are scored as X-items or *pure deviate* items and on another test are scored as O-items or *reverse deviate* items are separated by needle into two groups, an XO DEVIATE and an XO NON-DEVIATE group. The XO hole appears under the alignment cutout as shown in Figures 2 and 3. The items in the XO DEVIATE group are credited for X's, since in this group they are scored only as *pure deviate* (X) items. The items in the XO NON-DEVIATE group are credited for the O's, since in this group they are scored only as *reverse deviate* (O) items.

For illustrative purposes, the manner of notching the TRUE and FALSE holes for the three different types of items, *deviate* (X), *non-deviate* (O), and *mixed* (XO) is described in the Master Coding Chart and Table 1.

Table 1

Illustration of the Notching of TRUE and FALSE Holes for the Three Different Types of Items

Item	Type	Subtest Items	Cut Corner	Hole Punched
A-1	X	Hs, D, Hy	left	FALSE
A-12	X	Hs, Hy	right	TRUE
A-3	O	D, Hy	right	FALSE
B-6	O	D	left	TRUE
B-3	XO	D, Ma	right	TRUE
B-8	XO	Hs, D	left	FALSE

It can be seen that the *pure deviate* or X-items and the *mixed* or XO items have the TRUE hole notched where the cards are cut in the lower righthand corner, and have the FALSE hole notched where the cards are cut in the lower lefthand corner. The *reverse deviate* or O-items, on the other hand, have the TRUE and FALSE holes notched in exactly opposite fashion.

As for the subtests, the *pure deviate* (X) and the *reverse deviate* (O) cards have the holes notched for each subtest on which they appear. The *mixed* (XO) items have a double, or paired, set of holes around the edge of the card for each of the subtests, identified by the letters *d* and *n*. Where scored as a *deviate* or X-item on a given subtest, a notch is cut

through the *d* hole on the card. Where scored as a O or *non-deviate* item on a given subtest, a notch is cut through the *n* hole on the card. (See Figure 3.)

Procedures

The test is administered exactly as described in the MMPI Manual resulting in the same three stacks of cards: TRUE, FALSE, CANNOT SAY.

1. The CANNOT SAY cards are counted and entered in the ? space on the score sheet and then discarded.

2. The TRUE and FALSE stacks of cards are turned face down so that the needle enters through the blank side.

3. The needle is inserted through the TRUE hole on the TRUE stack of cards. The *deviate* cards drop out and are placed in a *deviate* group. The *non-deviate* cards remain on the needle and are placed in a *non-deviate* group.

4. The needle is inserted through the FALSE hole on the FALSE stack of cards. The *deviate* cards drop out and are added to the cards in the *deviate* group. The remaining cards are added to the cards in the *non-deviate* group. There are now two stacks of cards, *deviate* and *non-deviate*.

5. The needle is inserted through the XO hole in the *deviate* stack, releasing the *XO DEVIATE* cards, which are placed in a separate group. The cards remaining on the needle is the *deviate* stack.

6. The needle is inserted through the XO hole in the *non-deviate* stack, releasing the *XO NON-DEVIATE* cards, which are placed in a separate group. The *non-deviate* cards which remain on the needle are no longer used in scoring and are added to the CANNOT SAY cards which already have been discarded.

7. There are now three kinds of cards to score: (a) DEVIATES, (b) XO DEVIATES, (c) XO NON-DEVIATES. The cards are ready for the final scoring of the individual subtests.

8. The needle is inserted through the subtest holes of the DEVIATE group in counter-clockwise manner, beginning with the L hole. The cards which drop out are counted and *then returned to the DEVIATE stack*. Cards needled out of any stack, after being counted, are always returned to that stack prior to scoring the next subtest.

9. The twelve XO cards are scored in different fashion. For each of the subtests there is a double or paired set of holes around the edge of these cards. On the back or blank side of the card the letters *d* and *n* appear under the holes. Those cards which are scored as X or *deviate* items on a given subtest have the *d* hole notched. Those cards which

are scored as O or *NON-DEVIATE* items on a given subtest have the *n* hole notched. For each subtest in the XO *DEVIATE* group, the needle is inserted in the *d* hole and those cards which fall off the needle are counted. For each subtest in the XO *NON-DEVIATE* group, the

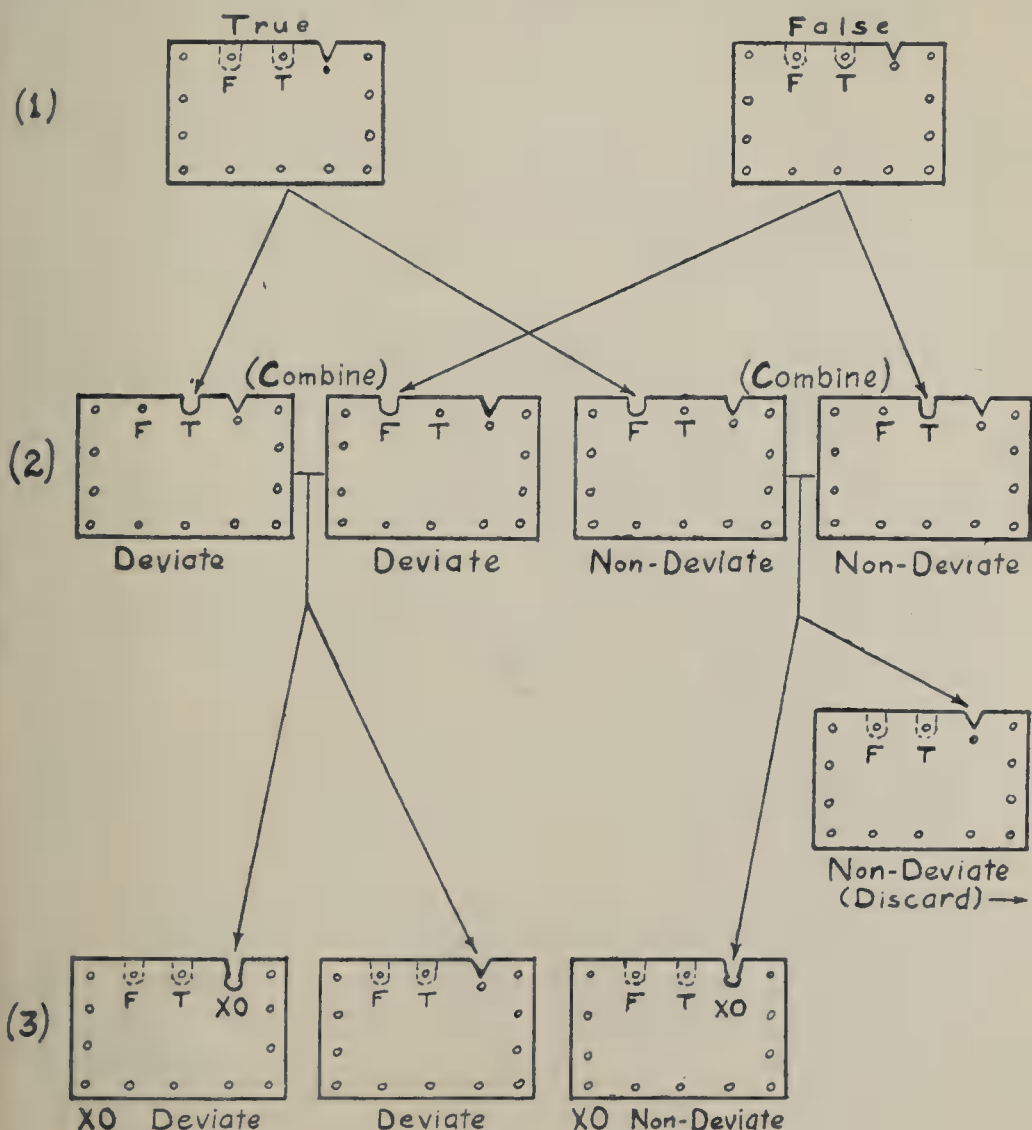


FIG. 4. Steps in breakdown.

needle is inserted in the *n* hole and similarly those cards which fall off the needle are counted.

10. The same procedure is gone through for each of the subtests. The score is always the sum of the number of cards dropped from the *DEVIATE*, the *XO DEVIATE*, and the *XO NON-DEVIATE* stacks.

In summary, the first needle movement, through the stack of TRUE cards, removes all cards which are placed in a DEVIATE group, the remainder being placed in a NON-DEVIATE group. The second needle movement, through the stack of FALSE cards, removes all cards which are added to the DEVIATE group, the remainder being added to the NON-DEVIATE group. The third needle movement, through the DEVIATE group, extracts the XO DEVIATE cards. The fourth needle movement, through the NON-DEVIATE group, extracts the XO NON-DEVIATE cards. The remaining stack is discarded. The subtest scores are obtained from the three stacks: DEVIATE, XO DEVIATE, XO NON-DEVIATE. (See Figure 4.)

The MMPI Manual lists twenty-six categories into which the cards have been classified and gives the number of items that appear in each category. It would be a simple matter to punch the cards, and code the holes, for the various categories, so that by the use of a needle the clinician could identify *deviate* and *non-deviate* answers in any area of special interest with regard to the particular patient.

Although the punch card method may appear to be complicated from the foregoing description, it is actually very simple in practice. One or two scoring attempts will demonstrate the simplicity, speed, and accuracy of this scoring method. The punch card technique described here can be adapted to the scoring of any tests which make use of, or can be converted to, a card system of administration.

The punch card system need not be limited to tests of personality. It can be applied equally well to any multiple-choice test or questionnaire. Test batteries of various types can be readily compiled and rapidly separated and scored by needling. Specific subtest areas, common to several tests, can be combined and analyzed. For example, all vocabulary items on several tests, or all spatial relations items, or all problem-solving items can be isolated, unified, and totally treated. New composite inter-test norms of increased reliability can be derived.

This method makes possible the rapid scoring of tests, enabling the clinician to devote more time to diagnosis and therapy. It enables him to use more tests and to make a quicker selection of specific items in interested areas of personality.

The elimination of paper and pencil, as provided by the MMPI, can be carried to the administration of many other tests. It now is practical for the examiner to reduce the administration of tests to a single instruction at the beginning, as in the MMPI, and to a rapid, simplified scoring at the end, as in this method of scoring.

Received November 9, 1945.

Profile Analysis of the Minnesota Multiphasic Personality Inventory in Differential Diagnosis *

Paul E. Meehl

University of Minnesota

A personality test may be employed in several kinds of clinical situations. These include: overall differentiation of normals from abnormals or persons predisposed to abnormal developments as in "screening" in the military, industrial, educational, or general medical out-patient situation; differential diagnosis among abnormals; prognosis; evaluation of changes and results of therapy; and the assessment of certain components for other than strictly diagnostic purposes such as the detection of important paranoid trends in a reactive depression even though diagnosis presents no problem.

The present paper presents preliminary data on the use of the Minnesota Multiphasic Personality Inventory (MMPI) with respect to differential diagnosis, with secondary findings upon the subject of overall identification of "abnormals" from people in general. Since MMPI has been described elsewhere (4), we may merely state that this device is a structured personality test which yields scores on nine components of abnormality, namely Hs (Hypochondriasis), D (Depression), Hy (Hysteria), Pd (Psychopathic deviate), Mf (Masculinity-femininity), Pa (Paranoia), Pt (Psychasthenia), Sc (Schizophrenia); and Ma (Hypomania). In addition, there are four scores which indicate "validity," in the sense that they attempt to detect test-records which for reasons such as confusion, language difficulty, or non-cooperation cannot be accepted as adequate samples of the patient's verbal behavior. These scores are ? (Cannot say), L ("Lie"), F (carelessness and misunderstanding), and a recently devised suppressor called K(5). For the details of function and interpretation of these validity indicators, the reader is referred to previous articles. In the present study, the scale Mf (Masculinity-femininity), has been excluded from consideration throughout, so that there are only eight personality components involved. All of the scores are expressed as T-scores, the general normal sample having a mean of 50 and a S.D. of 10.

* This article is a "prior publication," the author paying complete costs. The scheduled 80 pages per issue is thereby increased by the corresponding amount, thus the "early publication" of this article is a direct contribution to the subscribers of the *Journal of Applied Psychology* without handicap to those authors whose articles are accepted and printed in their regular turn.

The purpose of the present study was to evaluate MMPI as used in the differential diagnosis of three main categories of hospitalized psychiatric patients: psychosis, psychoneurosis, and "conduct disorder." Gough (2) and Schmidt (6) have stressed the importance of considering the "pattern" or configuration of the profile in addition to the elevation of single scores. An elevation on a single component, even if it is the highest or "peak" score of the profile, does not imply that the patient should be so diagnosed. For example, the most frequent peak score on abnormal profiles of all sorts is D (Depression). It is clinically known that many different kinds of psychiatric difficulties involve degrees of depression, and the test reflects this fact. Again, a peak of 75 on Sc might suggest a schizophrenic picture, whereas if it occurs together with markedly elevated scores on the neurotic triad (Hs, D, Hy) and a Pt of, say, 85, it may better be taken to indicate a psychoneurosis with poor prognosis (3). It must be emphasized that the patterning of a profile cannot be neglected in the case of structured tests any more than we would think of interpreting one determinant column of the Rorschach without considering anything else.

As yet, these configurational criteria on MMPI have not been adequately treated in the literature. Locally, the Minnesota group have tended to form more or less crude clinical judgments and global impressions based upon accumulated experience. The articles by Schmidt and Gough have contributed materially to the objectification of procedure, although neither of these investigators published results in the form of percent correct identifications for clinically diagnosed groups, a kind of treatment which is in many ways more meaningful than establishment of significant differences between central tendencies (1, p. 19). Furthermore, in both of these articles the similarity of "psychosis" to "severe psychoneurosis" in MMPI profile is too close for comfort, a drawback of MMPI which has been informally reported by a number of military clinical psychologists through personal communications.

In the present investigation, an attempt has been made to determine the approximate accuracy of a very rapid, inspectional diagnosis from the MMPI profile alone, using the more or less poorly defined criteria which have so far seemed valuable clinically. Naturally, it is not suggested that the profile be used in this way, but we want to know how much the test can contribute entirely on its own when so used. Because of the fact that recently hospitalized cases were not diagnosed independently of MMPI, it was necessary to utilize old cases, before July 1941, on whose response sheets the present scales had been subsequently scored. At that time, the MMPI had not been published and was still in process of development. The only scales which appeared on the profiles then in

use were H (a relatively less valid, uncorrected key for hypochondriasis) and D (Depression). For all practical purposes, it may be assumed that the clinical diagnoses made on these cases at that time were almost wholly unaffected by the presence of these scores on the chart. Of course, none of the present "pattern" criteria could have been employed at that time; further, knowledge of and confidence in the test were negligible among the psychiatric staff.

The procedure of blind diagnosis was as follows: profiles of male abnormals were leafed through in the order of their appearance in the files (roughly chronological). Any profile showing a ? (Cannot say) or L ("Lie") score as great as 70 was recorded as "invalid," except that if any abnormal score reached a standard score of 80, an elevated L score was ignored, since defensive lying could hardly be the reason for such a positive elevation. F was allowed to reach a raw score of 16 ($T = 80$) before the profile was considered invalid. The terms "valid" and "invalid" are used hereafter to indicate the acceptability of the profile as an adequate measure in terms of ?, L, and F, and have no reference to the question of accuracy of identification. When it had been decided that a profile was "valid" by these criteria, it was classified as either *normal* or *abnormal*. Actually, of course, it was known that all of the cases were abnormal, so that the criteria of classification had to be made wholly objective and hence more rigid than would be the case in practice. Profiles were called abnormal under the following four conditions:

1. Any of the eight components showed $T \geq 90$.
2. Any of the eight components showed $T \geq 80$, *unless* $K < 40$.¹
3. Any of the eight components showed $T \geq 70$, *unless* $K < 50$ and $L < 60$.
4. Any of the eight components showed $T \geq 65$, *unless* $K < 65$ and $L < 60$.

It can be seen from the above criteria that the classification into normal or abnormal is a matter of spotting the highest T-score, then reading to the right to see if the restrictions on K and L throw the profile into one group or the other. The profiles consist wholly of MMPI scores and a code number, so that there is no other source of information in making the diagnosis.

¹The scale K is a correction scale or suppressor variable which may be used to correct for certain test-taking attitudes which tend to invalidate a record. If the K score is low, it indicates that the testee was overly self-critical and obtained spuriously abnormal scores, hence is probably less deviate than his profile suggests. If K is high, it indicates a defensive tendency, and suggests that the profile is too low—a more subtle form of the old L scale. For further discussion see (5).

Application of these criteria to 294 profiles from our general population male sample² yields 10% "invalid" records on the basis of T, L, and F scores set as above. Of the records which can be accepted as valid, there are 9% indicative of abnormality by the criteria, which may be considered the upper limit of "false positives." Actually, of course, an unknown proportion of these false positives are profiles of persons who, although not under psychiatric care at the time of testing, were at least as psychiatrically deviant as some of the hospitalized abnormals. The figure 9% is to be contrasted with the 3% to 5% found previously for single scales. It is to be kept in mind in what follows that the differentiations achieved among the hospitalized abnormals occur at the expense of almost 1 in 10 among the normal population. The remainder of this paper deals only with the differentiation among actual abnormals.

When a profile had been classified as abnormal by the above criteria, a quick inspectional classification was made using three categories. The three employed were psychosis, psychoneurosis, and "conduct disorder." The last category is used to cover cases diagnosed constitutional psychopathic inferior, psychopathic personality, criminalism, alcoholism, except psychoses or deterioration, simple adult maladjustment, or "primary behavior disorder" such as the adolescent conduct problems not otherwise classified. The criteria employed in this subdivision of abnormal records were intentionally vague and subjective, since it was this sort of inspectional judgment which was to be evaluated. No "computations" of any sort were performed on the scores. In general, the criteria insofar as they were explicit, were those described by Schmidt and Gough, and such personal impressions as the examiner had acquired from considerable clinical work with MMPI. Psychosis was suggested by markedly elevated profiles, high F, Sc greater than Pt, Pa or Ma markedly elevated, the "psychotic" (right-hand) end of the curve reaching the level of the "neurotic" (left-hand) end, or a distinct spike on D, with the Hs and Hy scores on either side falling far below the D. Psychoneurosis was suggested by a less elevated profile, lower F, Pt greater than Sc, Pa and Ma not much elevated, the neurotic triad clearly elevated more than the rest of the curve, and the three scores of the triad closer to one another. Conduct disorder was suggested by elevations on Pd, Ma if not too high and especially with secondary peak at Pd, neurotic triad low except for some Hy, psychotic end running about 60. The examiner restricted himself to 10 seconds per profile in making his decision, and in most cases the judgment was made in less than five seconds. After having made the classifications, these were compared with the diagnoses of

² The data are for males only.

the psychiatric staff. All cases were eliminated in which the staff diagnosis was indicated as highly questionable or based upon insufficient study or cases of organic C.N.S. disease or feeble-mindedness. The actual composition of the abnormal group as subsequently determined was as follows: Psychosis, 57 cases (Schizophrenia 26, Manic-depressive 21, Paranoid condition 8, and Involutional melancholia 2); Psychoneurosis, 53 cases (Hypochondriasis 14, Hysteria 13, Reactive depression 9, Psychasthenia 7, Anxiety state 5, Mixed or unspecified 4, Neurasthenia 1); and Conduct Disorder, 37 cases (Psychopathic personality 21, Psychopathic personality pathological sexuality 8, Alcoholic 5, Behavior disorder 2, Adult maladjustment 1).

Of this entire group of 147 clinical abnormals, 25 (17%) invalidated their records on the basis of the validity indicators. Seventy-eight (53%) were correctly called abnormal, while the remaining 44 cases (30%) were (erroneously) classified as normal. The following table represents the data in various convenient breakdowns:

Table 1

Classification by Profile Inspection of 147 Records of Hospitalized Abnormals (Male)

A. Percentages based upon all 147 records

	Called Abnormal	Called Normal	Invalid Record
Total group ($N = 147$)	53%	30%	17%
Psychotics ($N = 57$)	60%	21%	19%
Neurotics ($N = 53$)	47%	36%	17%
Conduct disorder ($N = 37$)	51%	35%	14%

B. Percentages based upon the 122 valid records (based on T, L, F scores)

	Called Abnormal	Called Normal
Total group ($N = 122$)	64%	36%
Psychotics ($N = 46$)	74%	26%
Neurotics ($N = 44$)	57%	43%
Conduct disorder ($N = 32$)	59%	41%

C. Percentages based upon the 78 cases called abnormal

	Called Psychotic	Called Neurotic	Called Conduct Disorder
Psychotics ($N = 34$)	56%	29%	15%
Neurotics ($N = 25$)	24%	68%	8%
Conduct disorders ($N = 19$)	16%	16%	68%

From these tables, we see that in employing a criterion of abnormality which holds our false positives down to one in ten among general normals, we are able to detect only about half of the known abnormals (Table 1, A). This figure is not quite fair to the test, however, since those with invalid records would *not* under these conditions be erroneously classified as normal, but would either be requested to take the test over again with more precautions or their profiles disregarded. If we confine our attention to records in which the validity indicators are satisfactory, we find that about two-thirds of the abnormals can be identified (Table 1, B). It should be pointed out that the disappointingly large proportion of apparently invalid records among the abnormals (about one-sixth of all the records) is in part due to the time at which these tests were administered. At that time patients were allowed to invalidate their testings by sorting large numbers of the cards into the "Cannot say" category, sorting the cards at random, and so on. More systematic supervision now eliminates many of these uninterpretable profiles.

Setting up a contingency table for the 78 cases correctly classed as abnormal, we obtain a chi-square of 34.016, which with 4 d.f. is highly significant ($P < .001$). This corresponds to a contingency coefficient of .55, with the upper limit possible for a 3×3 table being .82.

In comparing the accuracy of identification for the three diagnostic groups, we shall consider only the valid testings, since the percentages of invalid records differ insignificantly among the three. While three-fourths of the psychotics were identified as being abnormal as contrasted with between one-half and three-fifths of the neurotics, a test of significance in proportion of false "normals" in the three diagnostic categories fails to show a significant difference (Chi-square 3.233, 2 d.f., $P > .14$). This being the case, most of the further subgroup differences in identification were not statistically analyzed. Mere inspection of Table 1, C, however, would suggest that the chief confusion occurs between neurotic and psychotic curves, rather than between either of these and the class of conduct disorders. Once having correctly classed a profile as being abnormal, the probability of its being thrown into the appropriate one of the three categories is about two in three.

Detailed inspection of the table of actual-real classifications does not indicate much because of the small numbers of cases in various subcategories. In the case of the psychoneuroses, however, inspection suggests that some clinical subgroups are more likely to show apparently "normal" profiles than are others. The differences in proportion called abnormal were tested by grouping the cases into four classes: hypochondriasis, hysteria, psychasthenia, and all others, and running a chi-square test on the resulting 4×2 table. This chi-square was barely significant

at the 5% level (Chi-square 8.303, 3 d.f., $P < .046$). Inspecting the table for the source of the differences, we find that 11 of the 13 hypochondriacs were identified as abnormal, as compared with only half of the ten hysterics, and only one of the six compulsives. It has been recognized for some time that the Pt scale is relatively ineffective clinically, and the use of K as a suppressor for Pt in this crude way tends to increase the false negatives by leading to an under-interpretation of profiles because K is highly correlated (negatively) with Pt. More detailed treatment of the actual subcategory tables is not warranted by the numbers involved.

Summary and Conclusions

The adequacy of the MMPI in differential diagnosis employing a rapid, inspectional method of pattern analysis of profiles was investigated by making "blind" diagnoses from records of 147 hospitalized psychiatric cases into three major categories of psychosis, psychoneurosis, and conduct disorder. The criterion was the clinical diagnosis of the psychiatric staff, made at a time when the present scales of MMPI, with one exception, were not yet in existence. The findings were as follows:

1. Setting up arbitrary criteria for the overall distinguishing of normal from abnormal persons, we find that about 1 in 10 persons from the general population sample is called abnormal (false positive).

2. Approximately $\frac{2}{3}$ of actual abnormal cases are identified as such by these criteria, if we exclude records obviously invalid on the basis of the validity indicators ψ , L, and F.

3. Of the abnormal cases identified as abnormal, about $\frac{2}{3}$ are placed in the appropriate category of the three employed. The contingency coefficient for the agreement between blind diagnostic grouping and the actual diagnosis is .55.

4. There is a suggestion that some varieties of abnormality are more readily identified than others. Hypochondriasis is fairly easily identified, whereas hysteria and psychasthenia are less so.

In general, while the discriminations achieved are very much better than chance in the statistical sense, especially considering the fact that *no* skilled clinical time is involved in giving or scoring the test and less than 10 seconds was used here in "interpreting" it, it must be admitted that the proportion of false classifications is considerable. Two developments can be expected to reduce materially this margin of error: first, the more mathematically precise utilization of the suppressor K; second, the greater formalization of pattern interpretation.

References

1. Dvorak, B. J. *Differential occupational ability patterns*. Employment Stabilization Research Institute Bulletin No. 8, Vol. 3. Minneapolis: University of Minnesota Press, 1935.
2. Gough, H. G. Diagnostic patterns on the Minnesota multiphasic personality inventory. *J. clin. Psychol.*, 1946, **2**, 23-37.
3. Harris, R. E., and Christiansen, C. Prediction of response to brief psychotherapy. *J. Psychol.*, 1946, **21**, 269-284.
4. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota multiphasic personality inventory*. New York: The Psychological Corporation, 1943.
5. Meehl, P. E., and Hathaway, S. R. The K factor as a suppressor variable in the Minnesota multiphasic personality inventory. *J. appl. Psychol.*, 1946, **30**, 525-564.
6. Schmidt, H. O. Test profiles as a diagnostic aid: The Minnesota multiphasic personality inventory. *J. appl. Psychol.*, 1945, **29**, 115-131.

The K Factor as a Suppressor Variable in the Minnesota Multiphasic Personality Inventory * †

Paul E. Meehl and Starke R. Hathaway

Division of Psychiatry and the Department of Psychology, University of Minnesota

I. History and Problem

Among the very large number of structured personality inventories which have been published, it is by now quite generally admitted that there are relatively few which are of practical value in the clinical situation. There are a number of reasons, both obvious and subtle, for this fact, some of which will be developed by implication in the present paper. One of the most important failings of almost all structured personality tests is their susceptibility to "faking" or "lying" in one way or another, as well as their even greater susceptibility to unconscious self-deception and role-playing on the part of individuals who may be consciously quite honest and sincere in their responses. The possibility of such factors having an invalidating effect upon the scores obtained has been mentioned by many writers, including Adams (1), Allport (2) (3) (4), Bernreuter (7) (8) (9), Bills (10), Bordin (11), Eisenberg and Wesman (15), Guilford and Guilford (18), Humm and Humm (31), Humm and Wadsworth (29), Kelly, Miles and Terman (32), Laird (33), Landis and Katz (34), Maller (39), Olson (51), Rosenzweig (53) (54), Ruch (55), Strong (58), Symonds (59), Vernon (62), Washburne (63), Willoughby (66) and others. One of the assumed advantages of the projective methods is that they are relatively less influenced by such distorting factors, although this assumption should be critically evaluated.

The existence of a distorting influence in test taking attitude is so obvious that it has hardly been thought necessary to establish it experimentally, although a number of investigations have demonstrated the effect. Frenkel-Brunswik (16) investigated tendencies to self-deception in rating oneself, finding in some cases marked negative relations between

* Supported by a research grant from the Graduate School of the University of Minnesota.

† This article is a "prior publication," the author paying complete costs. The scheduled 80 pages per issue is thereby increased by the corresponding amount, thus the "early publication" of this article is a direct contribution to the subscribers of the *Journal of Applied Psychology* without handicap to those authors whose articles are accepted and printed in their regular turn.

self-judgments and the evaluation of others. Hendrickson (27), cited by Olson (51), reported that a group of teachers earned significantly more stable, dominant, extroverted and self-sufficient scores on the Bernreuter scales when instructed to take the test as though they were applying for a position, than when under more neutral instructions. Ruch (55) showed that college students could fake extroversion on the Bernreuter to the extent of achieving a median at the 98th percentile of Bernreuter's norms, as contrasted with a "naive" median at the 50th percentile. Bernreuter (8) found that college students could produce marked shifts in their Bernreuter scores in the "socially approved" direction, although he interpreted this finding as indicating the comparative unimportance of the faking tendency. His reasoning was that had the need for giving socially approved responses operated in the first administration to any appreciable extent, the effect of special instructions to take this attitude should not have been great. This reasoning seems rather tenuous, inasmuch as the occurrence of a shift merely shows that conscious and permitted faking can produce greater effects than those which may have been operating in the "naive" original testing. The insignificant correlations between naive and faked scores were also used by Bernreuter to support his view, an argument which is not comprehensible to the present writers, especially in view of the probably gross skewness of the faked scores. What is clear from his investigation is that people are able to influence their scores to a considerable extent if they choose to, and that the average student's stereotype of what is "socially desirable" seems to be an individual who is dominant, self-sufficient, and stable. Maller (39), Metfessel (49), Olson (51) and Spencer (56) have studied the effects of anonymity on responses to self-rating situations and shown that the requirement of signing one's name has a definite effect on the scores. Kelly, Miles and Terman (32) demonstrated the great ease with which scores on the Terman-Miles Masculinity-Femininity Test could be "faked" in either direction once the subjects had been let in on the secret of what the test measured. Strong (58), Bills (10), Steinmetz (57), and Bordin (11) have presented evidence of the ability of subjects to distort their interest patterns when taking the Strong Vocational Interest Blank.

It is a significant sociological fact about the psychologists that in spite of the strong reasons, both a priori and experimental, for accepting the reality of this phenomenon in objective personality testing, very few systematic efforts have been made to correct for it or to overcome it. In published articles one continually finds brief and inadequate references to the "assumption of frankness" and the necessity for arousing a "sincere desire to know oneself better," but the treatment is usually extremely sketchy and no very concrete suggestions are given for producing such

test-taking attitudes nor, what is almost as important in practice, for determining the extent to which they have been present. It almost seems as though we inventory-makers were afraid to say too much about the problem because we had no effective solution for it, but it was too obvious a fact to be ignored so it was met by a polite nod. Meanwhile the scores obtained are subjected to varied and "precise" statistical manipulations which impel the student of behavior to wonder whether it is not the aim of the personality testers to get as far away from any unsanitary contact with the organism as possible. Part of this trend no doubt reflects the lack of clinical experience of some psychologists who concern themselves with personality testing, and the very strong contemporary trend which stresses the statistical interrelationships of item responses much more than the relation of the latter to external non-test criteria. The establishment of "validity" (sic!) in terms of various criteria of internal consistency naturally leads to an unconscious neglect of the problem of non-test behavior correlates.

Among the many authors who recognize the problem there are a few who have made specific suggestions for its solution. The inclusion of special exhortations to frankness and objectivity in the test directions themselves is common, but we have no evidence as to its effectiveness. Obviously, if a subject is consciously determined to fake, he will do so; whereas if his motivation to distortion is of a more subtle, non-verbalized nature, such exhortations can hardly be expected to be efficacious. Another method is to attempt disguise of the items, so that the "significance" of a given response is less obvious. Traditional approaches to the measurement of personality render this technique practically impossible, inasmuch as the items are selected to begin with for their *obvious* psychological significance and hence unless changed so greatly as to no longer elicit the desired information, almost inevitably continue to betray their origin. An effective use of a set of "subtle" items is only possible when the initial item pool is very large and the *initial selection* (not only the final validation) of items is ruthlessly empirical. Those items whose significance would not have been guessed by the test-maker will then be equally mysterious to the testee. When the projective and role-playing components of test-taking behavior are clearly seen to be present in objective personality inventories (46), this approach to the problem is very fruitful. A simple strategem along the item-disguise line is to state about half of the items negatively, so that an affirmative response is not consistently a "bad" or maladjusted one. However, such techniques cannot eliminate the problem entirely.

A spurious anonymity using secret coding for identifying the testee is a possibility suggested by the studies cited above, but is clinically

impractical for obvious reasons. The deception involved is not desirable, and in any case the clinical patient, unlike the sophomore student, knows perfectly well that the examiner is interested in *his* score individually. Lacking anonymity, it has been suggested by Olson (51) that the name be signed at the conclusion of the administration instead of at the top of the page. This suggestion was carried into practice by Maller (40) in his *Character sketches*. This investigator also stated the questions in the "indirect" (third person) form, requiring the subject to indicate whether he was the *same* or *different* from the person described. Maller presents evidence that this procedure aroused considerably less annoyance in his subjects, although direct proof that this decrease in annoyance led to increased validity is lacking. For reasons which have been given in more detail elsewhere (46), it is doubtful whether the removal of personal reference is wholly desirable; since there is reason for believing that the same role-playings and self-deceptions which operate to invalidate *some* of our measurements are an important factor in making *other* measurements possible.

Another technique for reducing the effect of signing one's name is to have the items printed on cards which are then sorted by the subject, making all writing unnecessary and possibly lessening the feeling that one is making a permanent record of his personal failings. This has been done by Maller in a revised test (Personality Sketches) and by Hathaway and McKinley in the Minnesota Multiphasic Personality Inventory (26). The latter test will be referred to as MMPI.

Although all of these stratagems may have a considerable value, especially in the aggregate, the fact still remains that they do not by any means remove the possibility of "faking." What is much more important, they are mainly directed at the sort of *conscious* falsehood which most writers have stressed, while ignoring the more subtle tendencies to self-deception which are probably of even greater importance in affecting scores. In the third place, they neglect to stress the existence of trends in the opposite direction—namely, those trends which exaggerate the apparent abnormality or maladjustment of the individual rather than soft-pedaling it. It is only natural that the tendency of a testee to put himself in a favorable light should have received more attention than the contrary tendency, which makes much less "sense" psychologically at least from a superficial point of view. There is evidence that this latter tendency does exist, however, and that it is a much more important factor in determining scores on personality inventories than has generally been supposed. Some of this evidence will be presented in the present paper, while other indications have been given elsewhere (47). It is also probable that certain systematic differences in item-interpretation, not

necessarily a function of personality dynamics of the defensive or self-critical sort but relatively "neutral" psychologically (e.g. semantic variation), lead to score deviations that are misleading. Such problems have been investigated by Benton (6), Eisenberg (14), and Eisenberg and Wesman (15).

A more fruitful attitude was taken by Rosenzweig (53) in which he reiterated the fact of untrustworthiness of self-ratings and indicated that instead of trying to completely eliminate these sources of error we should recognize them and attempt to "correct" for them in interpreting the results. He says,

"Astute phraseology in the instructions and questions of the test have sometimes been resorted to, but such expedients are rarely very effective. Might it not be more effective to recognize at the outset that such tests have certain limitations that can never be completely circumvented and then go on to the measurement of these limiting factors themselves, thus obtaining information by which a correction may be applied to the subject's answers?" (53).

Rosenzweig's specific proposal for achieving this end was to include among the usual self-rating items a set of items of the form "I should like to be the sort of man who . . ." on the theory that if we knew something of the strength of certain "ideal-self" trends in the person, we could make appropriate correction for these trends in interpreting responses to the traditional items. Rosenzweig never carried this idea into practice and there is no way of telling whether or not it would have worked. It seems to the writers that it would be relatively ineffective, since what is desired is not a statement of the strength or number of ideals for the self, but a measure of the extent to which they are allowed to distort responses. In other words, a subject might easily have quite lofty ideals verbally expressed, but might be too honest, insightful, objective, or self-critical to distort his responses into agreement with these ideals. It is, for example, rather characteristic of psychasthenic persons to express high and often unattainable ideals of perfection and achievement; whereas at the same time they are prone to be excessively self-critical, a fact which is psychometrically reflected in the negative correlation of the Pt (psychasthenia) scale of MMPI with some of the subtle "lie" scales which will be discussed below.

Maller (40) attempted to solve this problem in another way in his *Character sketches*, by including a small set of items which were supposed to measure the subject's "readiness to confide." The occurrence of very normal, well-adjusted scores in combination with a low measured "readiness to confide" would lead one to be sceptical of the validity of the measurement. This was a material advance in principle, except that the "readiness to confide" items were themselves self-ratings on that very readiness. In the later form called *Personality Sketches* Maller does not

make use of this procedure so we may assume that it was unsuccessful or at least did not materially improve validity.

Carrying Rosenzweig's thinking to its logical conclusion, the obvious procedure to follow is to give the subject a good *chance* to distort his answers in accordance with some self-picture or conscious façade, and observe the extent to which he does so. The difficulty here is that such a procedure requires a knowledge of the objective facts (and the subjective facts!) which is usually inaccessible to us. Here there are three possibilities open to the test-builder. First, he may sidestep the problem of getting directly at the objective truth, and attempt to establish falsehood by obtaining internal contradictions. This was another technique employed by Maller in his earlier test. Cady (13), in his application of a modified form of the Woodworth Psychoneurotic Inventory to the measurement of juvenile incorrigibility, had earlier made use of repeated items to increase reliability of the scores; although the aim of detecting inconsistency of the "fake" sort was not explicit in his rationale. Each question appeared twice, once in each section of the test, except that in the second appearance the question was phrased in the negative. Theoretically the subject's response should also be reversed; and the number of failures to reverse is an indication of some inconsistency and hence, Maller assumes of non-cooperation or dishonesty. The "inconsistency score" obtained in this way was to be subtracted from the adjustment score to get a sort of corrected score as proposed by Rosenzweig. It is by no means obvious that the shift to a negative form of item will leave the projective properties of the stimulus simply reversed in meaning; so that the fact of an "inconsistency" in the strict logical sense would not necessarily imply lack of cooperation or dishonesty. However, it would seem reasonable that a very large number of such inconsistent pairs would cast grave suspicion upon the scores, either for dishonesty or some equally serious reason. This technique also was abandoned by Maller in his revised instrument.

The second method of using distortion is to present opportunities for answering in a very favorable way but in a way which could almost certainly not be true. This idea was employed by Hartshorne and May in the Character Education Inquiry (23). Since there are very few aspects of behavior for which one could have complete confidence that no subject would be "ideal" in them, it is necessary to present a considerable number of such opportunities and progressively reduce the probability that any flesh-and-blood individual would be as described. Everyone has at least a few highly desirable traits, and no one has all of them. Without knowing anything whatsoever about a particular person, we can write down on common-sense grounds a list of extremely good and rare human qualities which it is statistically absurd to suppose will all or in

large part be his. If he says, however, that he has all (or a very great many) of them, we decide that he is not telling the truth. To practically clinch this argument it is only needful to choose desirable attributes which will very rarely belong, even singly, to anyone; and which furthermore relatively few normal persons claim for themselves when given the chance. In the mass the answers to these items may yield very strong evidence for deception. "I sometimes put off until tomorrow what I ought to do today" can be answered *False* by *very* few honest people. If a subject gives such responses with some considerable frequency, the inference is obvious. A more detailed discussion of this approach will be given in section III below.

The Humm-Wadsworth Temperament Scales and the Minnesota Multiphasic Personality Inventory have both made use of this method, the latter more explicitly. Humm and Wadsworth (29) deserve credit for having been among the first investigators of structured personality measurement to lay great stress upon the problem of detecting non-cooperation and distortion of response when evaluating a particular profile of scores. They were also among the first to adopt an explicit and uncompromising empiricism in selecting items from a large initial pool. The two scales which serve as "checks" or "correctors" for the remainder of the profile on the Humm-Wadsworth are the "Normal" component and the "no-count." The Normal component is rather difficult to evaluate from the theoretical point of view, for reasons which have been given elsewhere by one of the present writers (47). It is sufficient here to indicate merely its function as described by Humm and Wadsworth, which is to assess the strength of a general inhibiting, controlling, or normalizing factor in personality which serves to act as a "brake" upon strong abnormal tendencies on the other variables. This means that in interpreting a given profile, the significance of any deviation on one of the abnormal components must be established with the size of the Normal score in mind. To the extent that the Normal component measures what the authors claim for it, it is not especially relevant to the present problem; but if it actually operates by detecting something other than the personality component they describe, it would perhaps be of significance here. For a more detailed discussion of this question the reader is referred to the study cited above.

The "no-count" is based upon the number of items to which the subject responds in the negative. Inasmuch as approximately 76 per cent of the scored items (87 per cent of the total pool) of the Humm-Wadsworth are "obviously" suggestive of abnormality when replied to affirmatively, the "no-count" is to some extent a measure of the testee's tendency to avoid, consciously or otherwise, saying "bad" things about himself when

taking the test. That this relationship obtains is further supported by the tendency for the no-count to correlate positively (.77) with the "Normal" component and negatively ($-.39$ to $-.72$) with the various abnormal components (29). If the no-count is excessively great, the inference is that the subject has responded in a very defensive or possibly (as in some psychotics) stereotyped fashion; and therefore the particular testing is of doubtful validity. In another article, Humm and Wadsworth state that as high as 25 or 30 per cent of normals seem to invalidate their scores in this way, a proportion which would seem to be impractically high for clinical purposes. In a later article (30) they attempt to reduce the proportion of useless tests by a "correction" for the no-count based upon multiple regression procedures. Humm and Wadsworth state that in a subsequent group of cases "well known" to them, the improved validity of profiles thus corrected was demonstrated. An unpublished study of hospitalized psychiatric cases by Arnold (5) indicated that even the exclusion of cases with "invalid" no-count did not result in any greater validity clinically than was obtained using all cases. Humm (personal communication) states that improved multiple regression techniques have resulted in a very marked reduction in the proportion of test misses and of uninterpretable profiles. These more recent data on the Humm-Wadsworth have not been published. On present evidence it is difficult to say to what extent the use of multiple regression technique was successful in improving validity.

Washburne, in revising his "Test of Social Adjustment" (OSPA), included a set of 21 items modeled after the "lie" items of Hartshorne and May and referred to the total score on this set as *objectivity*. This score was included to detect both lying and unintentional inaccuracy, and the author reports that interviews with people showing very low objectivity scores showed that "it was useless to question them." A very low objectivity score was said to invalidate the test as a whole, and a weighted objectivity score was included in the total score on the entire test (63).

Another application of the second method for detecting invalidity by identifying the presence of distortion was the "lie" scale (and its complement, F) of the MMPI, which will be discussed in detail in section III below.

The third technique available is the empirical derivation of a "fake" scale by making use of the item shifts obtained when persons take a test under normal "naive" conditions and then are retested with instructions to fake. This method has been used by Ruch to construct an "honesty" key for the Bernreuter. It is interesting that a procedure so logical and straight-forward, invented to solve a problem so obvious and insistent,

should have been employed for the first time over twenty years after the appearance of the first personality inventory. Ruch says:

"The argument is rather simple. If answers to items on a test like the Bernreuter can be faked at all, the chances are that some are easier to fake than others. Therefore, it should be possible to give each item a weight to represent the extent to which it can be faked by the average college student. This was done by tabulating the frequency of each answer to each question for the standard condition and for the influenced condition. These frequencies were converted into percentages, and an 'honesty' weight was assigned to each reply according to the magnitude of the critical ratio of the difference between the frequency of the reply in the honest and in the influenced condition" (55).

In applying this honesty scale to a new group he was able to show that all cases of "real" introverts would be detected in an attempt to make themselves appear extroverted on the test. There are a number of interesting problems presented by this method, such as the extent to which the key would work if the subjects were not under actual instructions to fake extrovert but were being more "subtle" and actually trying to deceive an examiner in a real life situation. Presumably the deviation toward dishonesty would not be as great under such circumstances. The use of the critical ratio as a basis for weighting items might also be open to some question. In any event, Ruch seems to have been the first investigator to attempt empirical derivation of a fake key for a question-answer personality inventory. The results of applying this procedure to work on MMPI will follow in the present article.

As was mentioned earlier, there is some evidence of a tendency in the opposite direction in taking personality tests. It is difficult to characterize such a tendency, especially since it may occur on several different bases. A patient in the hospital may for instance engage in a sort of "psychiatric malingering" for strictly conscious reasons, presenting a profile on a test such as MMPI which shows abnormalities out of all reasonable proportion to what is apparent from other considerations. Again, there may be somewhat general traits of verbal pessimism or self-deprecation which, while of some relevance personologically, act so as to systematically distort the results of personality measurement. We shall dichotomize the test-attitude continuum by the two opposed terms "defensiveness" and "plus-getting," not implying anything as to the degree of conscious, deliberate deception involved in either. The corresponding *extremes*, where such deliberate deception seems likely, we shall refer to as "faking good" and "faking bad" respectively. It is recognized that, like the defensive tendency, the "plus-getting" tendency may exist in all degrees from a mild self-criticality or merely objectivity to a deliberate, conscious attempt to make oneself look psychiatrically abnormal. Whether this represents simply the extreme of a continuum

with faking good at the opposite end, or an entirely new and different factor, we shall for the moment leave aside. In any case it would be desirable to develop a scale for detecting these tendencies to put oneself in a bad light when answering a personality inventory, so that allowance might be made in such cases in the light of a deviant score obtained on such a scale. The F scale of MMPI was not originally developed with this in mind, but subsequent evidence showed that it could be used in this way (see below). Presumably the two "correction" scales C_h (42) and C_d (25) which were found necessary in the early attempts to detect hypochondriasis and symptomatic depression were at least partially dependent upon the operation of such a plus-getting tendency.

A systematic investigation of the plus-getting tendency was attempted by one of the writers, which resulted in the development of a somewhat more generalized correction scale which was called N. The details of derivation and interpretation of this scale are reported elsewhere (47) and will not be repeated here. Suffice it to say that from a study of the item responses made by a group of presumably normal persons who showed abnormal MMPI profiles as contrasted with a group of clinically abnormal persons with matched profiles, a group of items was isolated which could be used to roughly quantify the plus-getting tendency. It was found that normal persons who show distinctly abnormal (maladjusted) profiles on the personality scales proper, tended to answer this selected set of N items in the "obviously" maladjusted direction, which was with few exceptions also the direction of response given by a minority of the unselected normal population. In other words, a person who is clinically normal in spite of having an abnormal profile shows a tendency to give statistically uncommon answers which are also "maladjusted" answers in the sense that by inspection they would be considered evidence of psychiatric involvement. For example, about 48 per cent of the unselected general population normals answer "True" to the item "A windstorm terrifies me." Yet we find that among those normals selected specifically for showing apparently *abnormal* profiles on the personality scales proper, about 62 per cent give an affirmative answer to this question. Persons having MMPI profiles no more deviant than these plus-getting normals but who are actually abnormal clinically, give an affirmative answer about 26 per cent of the time. Thus if a person shows an otherwise deviant profile but states that he is terrified by windstorms he stands a better chance of being clinically normal than one who gives the a priori more "normal" or "adjusted" response. Similar items on the N scale include such things as "I am afraid of fire," "I have a fear of water," "People often disappoint me," "I did not like school," and so on. Inspection of these items and an examination of the correlations between N and the

other MMPI scales led to a conviction that the N scale was actually detecting a diffuse plus-getting tendency of the sort described. It was further shown that either the inspectional or mechanical use of the N scale in order to under-interpret profiles having the plus-getting tendency led to a reduction in the number of false positives in identification of psychiatric cases. However, the N scale was rather long, and was also apparently loaded with genuine psychiatric factors which led to an undesirable under-interpretation of profiles belonging to grossly abnormal persons. It is therefore to be seen merely as a beginning attempt which was supplanted by K as will be described below.

II. MMPI Scale F

The MMPI variables F and L were not formally validated originally, but were presented on face validity, that is, we assumed their validity on a priori grounds. The F variable was composed of 64 items that were selected primarily because they were answered with a relatively low frequency in either the true or false direction by the main normal group; the scored direction of response is the one which is rarely made by unselected normals. Additionally, the items were chosen to include a variety of content so that it was unlikely that any particular pattern would cause an individual to answer many of the items in the unusual direction. A few examples are: "Everything tastes the same." True. "I believe in law enforcement." False. "I see things, animals, or people around me that others do not see." True. The relative success of this selection of items, with the deliberate intent of forcing the average number of items answered in an unusual direction downward, is illustrated in the fact that the mean score on the 64 items runs between two and four points for all normal groups. The distribution curve is, of course, very skewed positively; and the higher scores approach half the number of items. In distributions of ordinary persons the frequency of scores drops very rapidly at about seven and is at the two or three per cent level by score twelve. Because of this quick cutting off of the curve the scores seven and twelve were arbitrarily assigned T-scored values of 60 and 70 in the original F table.

From the first it was recognized that F represented several things. Most simply, since the subject would need to sort almost all of the items according to expectation in order for these low scores to result, any error in recording, such as mistaking true items for false items and the like, would raise the F score appreciably. Similarly, if a subject could not understand what he was reading adequately enough to make conventional answers to these items, the F score would obviously be higher. It was felt to be axiomatic that this method would eliminate as invalid records

of subjects who could not read and comprehend or who refused to cooperate sufficiently to make expected placements.

In addition, however, it was early discovered that schizoid subjects and subjects who apparently wished to put themselves in a bad light also obtained high scores. The schizoid group obtained high scores because, due to delusional or other aberrant mental states, they said very unusual things in responding to the items and thus obtained high F scores. This is referred to as distortion since we feel that an impartial study would not justify the patient's placements. Among more normal persons some high scores were also observed where the individual had rather unusual ways of responding to conventional stimuli such as are represented by the items involved. For example, to the item, "I have had periods in which I carried on activities without knowing later what I had been doing," most persons answered false. Some persons, however, included periods of sleep in the implication of the item. One might argue that such ways of thinking are often allied to schizoid mentation generally and that the answers in this case indicate a true abnormality. At the very least, however, the person is responding to some items in a way that differs from that of most individuals. Such persons might, therefore, not be appropriately approached through this method of personality measurement. It seems a reasonable enough possibility that there are individuals whose habitual ways of reacting to items are so different from their fellows that measurement of their personalities through the use of verbal items of this type would reflect the unusualness of their reactions to the items more than any clinical abnormality. This semantic factor has been treated more completely elsewhere (6) (14) (47). In so far as such a possibility may exist we have not yet separated it from the clinically more important abnormality expressed in the Sc scale. Parenthetically, one of the most persistent difficulties with developing the Sc scale was this very fact, that an appreciable number of individuals obtained high scores on Sc without being marked by a clinically important degree of abnormality. They, nevertheless, as indicated above, were responding differently from other people about them as represented by the original data from the general population. It appears that the essential difference clinically is concerned with the particular manifestation of unusual mentation in the individual. If this is not too clearly counter to society's mores, the person may not be thought of as schizoid by those about him though he is often recognized as queer.

Clinical experience suggests that the usual critical score of $T = 70$ is too low in the case of F. We have found that scores ranging up to $T = 80$ (raw score 16) are more often a reflection of "validly" unusual symptoms and attitudes than an indication of invalidity in the rest of the

profile due to misunderstanding, etc. Raw scores much above this, however, strongly suggest an invalid record.

With the problems of measurement that developed in the armed forces where a subject might be expected frequently to attempt to put himself in a bad light in answering the MMPI, the F score became especially interesting. It was, of course, immediately possible to consider the F score as an evidence of this attempt to malingering and obtain fallaciously bad scores on other scales. As a check on this interpretation, a group of 54 ASTP men who had completed a considerable portion of their training in psychology were asked to take the Multiphasic twice.¹ The group took it in the usual way and also took it under instructions to assume that they wished to avoid being accepted in the draft and in order to be rejected they were to obtain adverse scores without giving themselves away. Half the group took it with "fake" instructions first, half second. Through this plan we had a "faked bad" record and a "normal" record. The data showed that 96 per cent of the "faked bad" records had a raw score F of 15 or more ($T \geq 78$), indicating that even these men who were somewhat cognizant of psychological measurements betrayed themselves when they attempted to fake a bad record. The resulting profiles were nearly all unusually abnormal in character, and most of them would have appeared invalid from their pattern alone to one who has seen a large number of abnormal profiles.

From this experiment it appeared that F was a very good device for identifying the intentional faking that could be set up in an experimental situation. It still seemed desirable, however, to attempt to separate among the individuals obtaining relatively high F scores those who were of the above-described schizoid type or who were simply over pessimistic in their view of themselves, from those who for one reason or another faked a bad score or did not understand the items.

III. The L Scale

The original 15 L items making up the L scale were selected under the inspiration of the work of Hartshorne and May. Each of the items presents a situation desirable socially, but rarely true of the individual. For example in the item, "I read all the editorials in the newspaper every day," we have a type of behavior that might be generally considered desirable but which is true of only a small percentage of the population. It was recognized that very conscientious persons would frequently have more than average of these L items truly positive but for a person to have six or eight of them seemed almost impossibly good. The 15 items of

¹ We are indebted to Dr. Howard F. Hunt for administration of these tests.

this type scattered among the main body of the items, constituted a fairly subtle trap for anyone who wanted to give an unusually good impression of himself.

Among the various normal groups the mean score on the L items lies between three and five. As in the case of F the frequency curves are all skewed sharply in the positive direction. Very few individuals obtain raw scores of seven or more, and the two or three per cent level is at about ten. These values were arbitrarily called the 60 and 70 T-score points, respectively. As the L score was used in the clinical setting and as some data began to accumulate from personnel workers in industrial situations, it became apparent that the assumptions regarding the meaning of L were in the main correct, but that there were also other valid interpretations of L, at least in the range from T-score 56 to 70. In fact we found ourselves placing considerable emphasis on T-scores of 56 to 60 which indicated that the original arbitrary assignment of T-scores had been too conservative. On the other hand while the positive presence of the rise in the L score seemed quite valid as an indicator that the individual taking the test was being dishonest and might be somewhat unreliable, if no rise in L was observed, the finding could not be so positively and clearly interpreted. The L score was a trap for the naive subject but easily avoided by more sophisticated subjects.

To check the assumption that L would not identify the more sophisticated subject an experiment was performed with ASTP psychology students. As in the study cited under Section II above, 53 men were given the MMPI twice. The "faked good" data were obtained under the instruction to make certain in taking the test that they would be acceptable to army induction. These records showed no appreciable rise in L. It is also true, however, that the majority of the profiles were only slightly, if any, better than the corresponding non-fake profiles. This experiment would have been improved if persons whose true profiles were abnormal had been used. Some data have been collected from such cases but the number is small. At least, one may conclude that the intent to deceive is not often detectable by L when the subjects are relatively normal and sophisticated.

IV. The K Scale

In summary there were two basic lines of experimental approach to the problem of identifying the attitude a subject takes toward the items that he is faced with in the personality inventory.² Each of these two

² Harmon and Wiener (personal communication) have investigated the possibility of detecting defensive and plus-getting tendencies through a division of certain MMPI scales into "subtle" and "obvious" items. Separate T-scores may then be calculated

approaches permits a subdivision into several methods. First, we may have the subject deliberately assume a generally defined attitude, as in the study by Ruch. For example, we may ask him to attempt deliberately to obtain adverse scores while not betraying his intention, and secondly, we may choose records in which there is presumptive likelihood that a special attitude has been assumed. The first approach may be subdivided into those experiments in which the "faking" is directed toward obtaining adverse scores and the approach in which the intention is to obtain desirable scores. In both latter cases an additional set of responses must be obtained relatively simultaneously with the "faked" responses in which the individual assumes his ordinary attitude. The "faked" and "normal" records can then be contrasted for study. One may then make an item analysis to discover the items that are most frequently changed from the "normal" records as contrasted to the "fake" records. Using these "fake" approaches, several scales were derived.

It was found that the items indicating an attempt to obtain a bad record are not necessarily those derived by analysis of records where the subjects attempted to obtain a good record. Our first finding in this regard was that either of these procedures provided a scale that would be about as good for the other type of faking as it was for the one from which it was derived when such scales were applied to test cases not used in the original derivations. It was further found that using two such scales separately did not materially increase the predictive value. As has already been pointed out, it was also found that the original F scale was as effective as was needed to identify those persons who intentionally attempted to obtain a bad score at least within the range of the experiments that we conducted. Conversely, the L scale was not effective nor were any of the specially derived scales especially effective in identifying sophisticated persons who deliberately attempted to obtain better scores. In all of these experiments the findings were so complex and the time devoted to many subprojects was so great that we shall only present data for the final scale K (see below).

In the second line of experimental approach there are also several subdivisions. One may find among presumably functional and normal records those records which are so abnormal as to indicate that the individual should have been in a hospital and attempt to discover the items

for the subtle and obvious scores on each scale so treated, and in terms of the discrepancy between S and O one may form a judgment as to the strength of the defensive or plus-getting test attitude of the subject. This ingenious technique is still in process of investigation by its inventors and a more adequate treatment of the method and its results will presumably be forthcoming from them later.

among these records that will differentiate them from the records of actually abnormal persons. For the counterpart to this approach one chooses cases who were in the hospital but whose records show a normal profile. These may likewise be compared by item analysis to the records of hospital patients with suitably abnormal profiles who would be assumed to have had no interfering test taking attitude. Using this approach we also derived several scales and made many experimental tests of them. Again the details of all of these are not worthy of the complex presentation they would require and these preliminary results will merely be summarized.

The first and most important finding was that whichever of these methods was used, as was the case with the "faked" approach above, the resultant scales were about equally effective and about equally unsatisfactory regardless of the approach and of the particular item content. These scales were also rather effective in differentiating the "fake" group and in some cases were just as valid for that purpose as were the scales derived by that approach. After some two years of this experimentation all of the scales that had showed any promise were reconsidered by applying them to various available groups that had not been used in their derivation and from among them all a single scale which was originally called L6 was chosen as the best. It should be recognized that L6 was not entirely satisfactory but its action in several of the sample situations resulted in its tentative adoption. Although as indicated in the above summary the particular derivation does not seem to play an important part since we could not easily distinguish a scale as having been derived by a special process when we examined its action; nevertheless it may be desirable to tell how L6 was derived. It must not be forgotten that several other scales resulting from the other methods were very nearly as good as was L6, especially the plus-getting scale N. However, when the N scale and L6 were compared and even applied to the test situation set up for the N scale, L6 was a close competitor with N and in several instances was actually better.

In brief, L6 was derived by an item analysis of the responses of 25 males and 25 females in the psychopathic hospital whose profiles showed an L score of $T = 60$ or more and who, with the exception of six normal cases, had diagnoses indicating the probability that they should have had abnormal profiles but whose profiles were in reality within normal range. The diagnoses given to these cases by the psychiatric staff were mostly psychopathic personality, alcoholism and allied descriptive terms indicating behavior disorders rather than neuroses. In general one would expect persons with such diagnoses to be rather more likely to be defensive in taking a personality test than cases of psychoneurosis. There are a few

exceptions, however, in the case of hysteria where as has been pointed out in previous papers (44, 46, 47) there is a tendency for the hysteria to be based upon something closely allied to the assuming of an overly perfect attitude in answering personality items. A particular listing of the diagnoses among these cases is not given here because the diagnostic categories are not clear enough to be of additional value. In summary, two criteria were employed in the selection of the criterion group. Practically all of them were individuals known to be characterized by deviant behavior but they obtained relatively normal profiles and were thus what we have called misses for the Inventory; and all of these criterion cases were also characterized by having a tendency to obtain elevated scores on the original L scale.

The item responses of these fifty cases handled separately for males and females were compared to the male and female item frequencies from the general group of males and females that has been used in past scale derivations. In all, 22 items were chosen as a result of this comparison. All of these items showed a per cent difference of 30 or more between the criterion cases and the control group, males and females being considered separately.

These 22 items are given below with a T or an F to indicate the direction of the answer that would agree with the tendency of the criterion group. Since the criterion group is assumed to be a group with intent to obtain good scores the larger raw scores on these items are in the same direction as the larger raw scores on the original L scale. The content of these items would seem to suggest an attitude of denying worries, inferiority feelings, and psychiatrically unhealthy symptoms, together with a disposition to see only good in others as well as oneself. The over-all impression is one of "impunitiveness" if we may apply Rosenzweig's term in a situation where no particular frustration is involved.

- B-55 I have very few quarrels with members of my family. (T)
- C-18 I worry over money and business. (F)
- C-27 It makes me impatient to have people ask my advice or otherwise interrupt me when I am working on something important. (F)
- C-33 It takes a lot of argument to convince most people of the truth. (F)
- D-53 I think nearly anyone would tell a lie to keep out of trouble. (F)
- D-54 Most people will use somewhat unfair means to gain profit or an advantage rather than to lose it. (F)
- E-43 When in a group of people I have trouble thinking of the right things to talk about. (F)

- E-44 I find it hard to make talk when I meet new people. (F)
E-52 People often disappoint me. (F)
F-8 It makes me uncomfortable to put on a stunt at a party even when others are doing the same sort of things. (F)
F-33 Often I can't understand why I have been so cross and grouchy. (F)
F-34 Criticism or scolding hurts me terribly. (F)
F-43 At periods my mind seems to work more slowly than usual. (F)
F-46 I frequently find myself worrying about something. (F)
G-18 I have periods in which I feel unusually cheerful without any special reason. (F)
G-29 I get mad easily and then get over it soon. (F)
G-30 At times my thoughts have raced ahead faster than I could speak them. (F)
G-31 At times I feel like smashing things. (F)
I-22 I have often met people who were supposed to be experts who were no better than I. (F)
I-31 I have sometimes felt that difficulties were piling up so high that I could not overcome them. (F)
I-37 I certainly feel useless at times. (F)
I-38 I often think "I wish I were a child again." (F)

Following the final choice of L_6 as the best of the scales available, we subjected it to more careful study and went back through hospital and normal records to find out if it seemed to be of any help in interpreting individual profiles. There were relatively few data on normal cases but on hospital cases a fairly extensive symptomatic summary was available that would permit us to judge whether or not a patient should have had a normal profile. We could then look up the profile and if it was normal we could check to see if the L_6 deviated in an upward direction indicating that the patient had attempted to place himself in a good light. As a result of this study L_6 appeared effective but left much to be desired.

Since in the summary of scales when L_6 was chosen for intensive study, it had seemed about as adequate for the detection of plus-getting as was N or any of the other experimental scales, the records of a new series of presumably normal persons showing deviant profiles was examined and it was again true that L_6 appeared to work at the plus-getting end of the test-attitude continuum. That is to say, a relatively low score on L_6 could be used to under-interpret an otherwise deviant profile and thus avoid some of the presumably false positives in the normal population sample. Thus L_6 seemed useful at "both ends" of the test-attitude continuum, defensiveness and plus-getting.

The most outstanding difficulty in such a procedure was that L_6 tended to be low on severe depressive or schizophrenic patient records and thus lead to an under-interpretation in spite of the fact that the patients were very grossly abnormal. To partly correct for this tendency, items were added that would work in the opposite direction. To choose these we studied the item tabulations for the group of ASTP men who had attempted to fake good and bad scores. In this study there were many items which showed no tendency to change with an alteration in the test-taking attitude. That is, the per cent of true or false, as the case might be, remained constant whether the attitude was the normal one or the faked one. From among these items, a sub-group was chosen which showed differences between schizophrenic and depressive criterion groups and general population normals. The procedure rested upon the admittedly somewhat shaky assumption that any item that did not appear to be much affected by the test-taking attitude as approached by a normal person attempting consciously to "fake" good or bad but which did occur as a frequent item to differentiate depressed or schizophrenic patients would be useful in correcting the tendency of our L_6 scale to go too low for schizophrenic and depressed patients. Of course such an item was scored in a way that would make it work against the tendency of the L_6 scale. Eight items were selected by this method. The effect of adding these eight items to the 22 on L_6 was of course to elevate slightly the mean score of normals and make it more nearly approach the mean score of abnormal cases on the complex of all 30 items. The eight items chosen by this procedure are given below. The letters T and F indicate the response scored in the "lie" direction, and in the direction characteristic of schizophrenic and depressed cases.

- A-3 I have never felt better in my life than I do now. (F)
- C-28 I find it hard to set aside a task that I have undertaken, even for a short time. (F)
- D-48 I think a great many people exaggerate their misfortunes in order to gain the sympathy and help of others. (F)
- D-51 I am against giving money to beggars. (F)
- F-7 What others think of me does not bother me. (F)
- F-20 I like to let people know where I stand on things. (F)
- G-23 At times I am all full of energy. (F)
- J-51 At times I feel like swearing. (F)

As a final step these eight items were combined with the 22 L_6 items into a single scale which we have called K. The K scale represents the final outcome of many experiments in the general field of measuring test

attitude. The K scale is far from perfect for its purpose as measured by the various available data. Generally speaking it is about as good as any other single scale derived for any one of the single purposes that have been described. In individual applications it is inferior now to one scale and now to another but the differences are never great enough to be very significant practically and the small number of items in this scale gives it a distinct advantage over one or two of the longer scales such as N. Finally, as was stated above it is not expedient to present more than a single scale although a slight advantage could have been gained if two scales analogous to the original L and F scales had been separately presented.

The construction of K being what it was, odd-even or Kuder-Richardson reliabilities were not computed. Test-retest coefficients were .72 and .74 computed on two groups, one of which was retested at intervals varying from one day to over a year, the other after a lapse of 4-15 months.

Since the K scale was derived as a correction scale or suppressor variable (28, 48) for improving the discrimination yielded on the already existent personality scales, it was not assumed to be measuring anything which in itself is of psychiatric significance. Actually, its relationship with such clinical variables as the subtle Hy items (see below) might suggest an interpretation of K alone; further, it is presumably a significant fact about a person that, in answering a personality inventory, he tends to behave as a "liar" or a "plus-getter." However, the real function of K is intended to be the correction of the other scores; and validity will be discussed with reference to this function only.

It is first necessary to choose criterion cases of the sort on which K can conceivably be of value. It is clear that such cases will be characterized by the presence of what may be called *borderline* profiles, i.e., those showing T-scores, say, between 65 and 80. The reason for this is that in studying hundreds of deviant profiles after the addition of K, almost no individuals were found with T-scores above 80 in the normal sample, and it was not statistically profitable to correct elevations of such magnitude to the point of calling them normal. On the other hand, when a curve shows no elevations at all above 65, even the presence of a high K score does not enable the clinician to form any adequate notion of what the peak would be, if any, had the K-factor not been operating to distort the results. In other words, there are upper and lower limits beyond which deviations on K cannot effectively operate. Profiles showing scores above 80 are to be interpreted as probably "abnormal" no matter how low K falls; while if a profile shows no scores above 65 we cannot tell whether a high K means the profile should be adjusted toward more severe scores or is merely that of an actually normal person who for some reason or other took a defensive attitude when being tested. The kind of curve

which gives interpretative difficulty and which could conceivably be improved by knowledge of the influence of *K* would be a curve in the doubtful, borderline region. Accordingly, a group of cases from the normal and hospital groups was chosen on the basis of having achieved such borderline curves. We selected for this study all cases in the files showing at least one personality component³ elevated as high as $T = 65$, but no component elevated to $T > 80$. Among the normals, there were 174 having such borderline curves, of which 71 were males and 103 were females. Corresponding to these cases, we located among our clinically abnormal cases 129 males and 208 females with similar borderline profiles. The data for the two sexes were treated separately.

The analysis of these data was in terms of the ability of the *K* scale, used mechanically as will be described, to separate the curves of the actual normals from those of the actual abnormals. For each sex group, the procedure was to arrange the whole set (normals and abnormals combined) in order of the magnitude of their *K* scores. The distribution of *K* was cut on the basis of the proportion of normals and abnormals in the sample, calling all cases above the cut "abnormal" and all those below "normal." Setting up a fourfold table on this basis, a chi-square of 20.436 for the males and 29.540 for the females was obtained. Both of these are highly significant ($P < .001$) with 1 d.f. If, instead of locating an optimal cutting score the *K* distribution was cut at the mean of the general population *K* distribution (i.e., at $T = 50$ regardless of the present samples) the cutting point of the males is unchanged, whereas that for the females shifts enough to lower their chi-square to 17.750, which is still highly significant. In other words, if one considers miscellaneous profiles which lie in the borderline range between 65 and 80, regardless of the kind of elevation and irrespective of the clinical diagnosis of those who are clinically abnormal, he can separate them into "actual" normals and abnormals significantly better than chance by using a cutting score on *K*. It must be emphasized again that *K* in this instance is operating chiefly as a suppressor of certain test-taking tendencies, since *K* by itself does not practically differentiate unselected normal and abnormal cases (1 to $2\frac{1}{2}$ raw score points difference between means for various samples). In terms of percentages, it was found that for the males, 72 per cent of the abnormals and 61 per cent of the actual normals were correctly identified. For the females, 66 per cent of the abnormals were identified as such and 59 per cent of the normals were so classified. These percentages are based upon the separations at a $K = 50$, taking, therefore, no account of the actual normal-abnormal proportions among the present cases.

³ Mf is excluded from consideration here and in all that follows.

Evidence from examination of the test misses spotted by K in the above data combined with our knowledge of the correlation between K and other MMPI scales, indicated that the K correction was more important in the case of some scales than of others. Therefore, it was decided to analyze the borderline groups in terms of the peak elevation of their profiles, in the attempt to identify those particular curves on which K could be used with profit.

The entire group of 511 borderline curves (males and females, normals and abnormals pooled) was divided into eight sub-groups, each sub-group being composed of cases having the peak score on the same one of the eight personality components. Thus, there were 60 curves having the peak on Hs, 91 on D, 119 on Hy, 66 on Pd, 38 on Pa, 25 on Pt, 28 on Sc, and 52 on Ma. (The difference between this total of 479 cases and the 511 used in getting the over-all chi-square is due to the exclusion of 32 profiles on which no "peak" could be fairly assigned, since two or more of the components showed identical T-scores and these were the highest on the given curve.)

The normals and abnormals having borderline curves with the same peak score were then separated mechanically by the use of a cutting score on K, the proportion of cases above the cutting score being determined on the basis of the proportion of actual abnormals versus normals in each sub-group. This was unavoidable in the present analysis because the relative proportions of actual normals and abnormals varied widely from scale to scale and the use of the mean of K would have been grossly misleading since in some instances the proportions were extremely asymmetrical (67). For the eight groups studied in this manner, only three showed a significant chi-square ($P < .01$), namely those having peaks on Hs, Pd, and Sc. The Ma group yielded a chi-square between the 10 per cent and 20 per cent level of significance. On D, Hy, Pa and Pt the chi-squares were all below the 20 per cent level of significance; and the pooled chi-square for these five scales (5 d.f.) gave a $P > .22$. It would seem, therefore, that the K-factor may be used with profit in interpreting some kinds of profiles but not others. Of course, the failure to discriminate with K when grouping profiles by peak score does not establish that a K-correction might not be profitably added to the single scores themselves. This problem will be treated at length in a sequel to the present paper.

One other validating study was done on K. In this instance, we made use of a group of 22 normals and 22 abnormals employed in a previous study (47). The normals in this set consisted of a random selection from a large group of profiles showing any elevation of 70 or over (excluding Mf). The abnormals consisted of a heterogeneous group also having at least one

score over 70, and included seven psychoneurotics, seven schizophrenics, three psychopaths, two alcoholics, two manic-depressive (depressed), and one paranoid state, chosen randomly from recent hospital cases. These groups had been selected for a different purpose and had not entered into the derivation of K in any way. They can also be considered, therefore, a fair test group for validation purposes. Without regard for any other information concerning the profiles, all cases showing $K > 50$ were arbitrarily guessed as abnormals, whereas those with $K < 50$ were called normals. The cutting score was therefore also independent of the statistics of the present group. Here the K scale worked phenomenally well, being much better than the N -scale (which was derived on cases some of which were included in this blind diagnosis study). Of the entire group of 44 cases, 37 were correctly classified when using K in this way, a total of 85 per cent hits. It will be recalled that we are here trying to separate normals and abnormals all of whom have deviant profiles, so that this per cent is quite impressive considering the task set for K . Of the seven errors in classifying, six are "false positives," i.e., cases of normals showing elevated profiles and $K > 50$, called therefore abnormal. The chi-square for the fourfold table of these data is 21.569 which with 1 d.f. is highly significant ($P < .001$). This corresponds to a contingency coefficient of .57. Here we have striking evidence of the validity of K when used to differentiate between deviant curves of actual normals and abnormals. We are not prepared to explain the superiority of this result to that given by the analysis previously discussed, except to say that the range of abnormal scores in the present analysis was from 70 to 90 whereas in the previous analysis we used "borderline" scores defined as lying between 65 and 80. In what way this could make K appear to function more effectively in the one case than the other is not clear. Also the present study involved only males, where K in general seems to work a little better than on females.

The fact that K is less effective as applied to some scales than others would suggest separate interpretations or cutting scores depending upon the kind of profile with which one is confronted. Furthermore, the rough classification into "normal" and "abnormal" on the basis of a single arbitrary cutting score obviously sacrifices some quantitative information about the actual magnitude of the personality scale elevations with respect to the magnitude of the K score. We do not intend to propose such a rough cutting method as the most efficient manner of application for K , but are using that form here simply to indicate that K has differentiating power for what it was hoped to differentiate. The optimal mathematical procedure in using K as a suppressor involves complex issues which we shall have to reserve for a later publication.

V. Relation of K to Other Test Variables

The correlation of the K scale with other MMPI variables should throw some light upon the question of its differential efficiency on these scales, as well as give us some insight into its psychological nature. Table 1 below shows the intercorrelations of K with the other personality components measured by MMPI. These correlations are based upon 100 cases in each of the four groups indicated, chronological ages 26-45, excluding records having "F" > 70 or *F* > 80.

Table 1
Intercorrelations of K with Other MMPI Variables

	Hs	D	Hy	Pd	Pa	Pt	Sc	Ma	Mf
Normal males	-.30	.15	.48	-.17	-.07	-.67	-.59	-.36	
Normal females	-.35	-.03	.30	-.06	-.02	-.64	-.58	-.28	
Male abnormals	-.42	-.29	.11	-.26	-.19	-.60	-.60	-.37	-.08
Female abnormals	-.17	-.16	.17	-.21	-.13	-.63	-.58	-.38	.04

Of interest in this table are the following facts. With the exception of Hy and one of the four coefficients of D, the correlations are consistently negative. This is of course to be expected if K represents the defensive, lying, or self-deceptive test-taking attitude it was derived to measure. The negative correlations with Hs combined with the positive correlation with Hy indicate that there must be a fairly high positive correlation between K and those non-somatic items on Hy which have been previously referred to—the "zero" items on Hy or what Harmon and Wiener have called "hy-subtle" (henceforth designated Hy-O).⁴ Since this latter set of items, although derived by its empirical separation of clinical hysterics from normals, seems to reflect the self-deceptive and impulsive attitude of the hysterical temperament, it is consonant with our interpretation of K that it should be markedly correlated with Hy-O. The direct evidence on this point will be reported below. The only correlations of very impressive magnitude which appear in this table are those with Pt and Sc. Here they are high negative—the person who makes responses characteristic of compulsive and schizoid persons has the opposite of the self-deceptive and defensive attitude. In other words, he tends to be a "plus-getter" and in this way is distinctly unlike the hysteric.

⁴ These items are called "zero" items because on the scoring templates they are indicated with a letter "O," meaning that one receives a point for the "abnormality" by responding in the direction which, on that single item, characterizes the majority of general normals. This means that the abnormals in question tend to give the "normal" response much more often than the normals do.

These correlations are also in harmony with our clinical knowledge of the components in question, especially in the case of the psychasthenia. The Pt scale has never been considered very satisfactory, and it has been shown in unpublished studies that Pt can actually be used as a correction scale in the way in which N was used. It is perhaps significant that of all the MMPI scales, Pt is the only one for which, lacking a sufficiently large criterion group, methods of internal consistency were employed in the item selection. Here again we would expect to get a greater operation of non-clinical test-taking factors of the K variety.

It might be thought that such low correlations as occur in the table above would preclude any possibility of the use of K as a suppressor. There is a tendency for the scales on which K seems "valid" by the chi-square test to show the higher correlations, with the exception of Pt. It will be shown in a subsequent paper that, for the use to which K is put, correlations as low as .20 can be utilized to yield very significant and useful improvements in discrimination.

At this point we may briefly review some of the previously developed scales which are now known to be saturated with what we may call the *K-factor*, since their diverse sources and methods of derivation furnish additional strong evidence for our theoretical interpretation of K. Two of these scales have never been published, so that their derivation and properties must be briefly summarized here. About three years before research on the test-taking attitude was begun, Hathaway and W. K. Estes, using a variant of the method of internal consistency, developed a scale called G. This scale is the only MMPI scale which was derived without the use of any kind of criterion external to the test; like those personality tests being developed by factor analytic methods at the present time, the selection and scoring of items was based wholly upon the intercorrelations among the items themselves. Essentially, the procedure consisted in locating among a group of 101 unselected normals those individuals who, when their answer sheets were used as scoring keys, produced the maximum variance of the other 100 scores. The assumption was that these persons were the most extreme deviates on whatever factor or factors contributed most heavily to the variance and covariance of the total pool of MMPI items. From the evidence adduced by Mosier (50), it is of course clear that the "purity" or factorial unity of this hypothetical underlying continuum is by no means guaranteed by such a procedure. Another way of looking at this procedure is to consider the fact that one maximizes the variance of a set of items by scoring them in such a direction as to maximize their mean covariance—since the item variances are unaffected by the direction of scoring. Instead of actually calculating the variances for the 2^{550} ways of scoring the test, we

select *individuals* who approximate the optimal scoring key. It was found that the scoring keys for some 10 individuals selected by this method tended to form two distinct clusters, each of which consisted of keys (individuals) showing high correlations with one another and high negative correlations with the members of the other cluster. An item analysis was then carried out on these two small groups, and the items resulting were combined into a scale called G (general factor).

The G scale had a number of interesting properties which were not interpretable at the time of its derivation. It showed a very large variability, both in absolute terms and as indicated by a coefficient of variation. The scores *among normals* ranged from those who answered none of the items in the scored direction, to those who answered all but eight of the 62 items in the scored direction—a phenomenon unheard of in the other MMPI scales. The odd-even reliability of G was about .93, which is considerably higher than the coefficients we typically find in the MMPI scales. The item content was that of the typical “neurotic” or “maladjustment” sort which predominates on *a priori* scales such as the Thurstone or Bernreuter BI-N. Examples of items are: “When in a group of people I have trouble thinking of the right things to talk about” (T); “I cry easily” (T); “I am certainly lacking in self confidence” (T). It is perhaps significant that the most powerful single item in the internal consistency sense—which happens in the sample studies to have a correlation of 1.00 with the entire G-scale—is almost a distilled essence or prototype of so-called “neurotic schedule” items: “I am easily embarrassed” (T). The G scale, although derived without recourse to any clinical group whatever, nevertheless showed a correlation of .91 with Pt. The mean MMPI curves for unselected normals with high G (the “neurotic” end) showed elevations on F, Hs, D, Pd, Pa, Pt, Sc, and Ma, especially on Pt and Sc; whereas L (raw score) and Hy tended to fall below the mean. The mean profile for normals with low G was almost an exact mirror image of this curve. However, G was not found to be very effective in the detection of any clinical group or to be particularly useful for any purpose; and since at that time no theoretical basis was available for interpreting it, the scale was abandoned. Another scale, called + (“plus”), was derived in a similar but not identical manner.

In the derivation of the original hypochondriasis key, there was developed a correction scale called Ch, the function of which was to separate actual clinical hypochondriacs from a group of non-hypochondriacal abnormals (mostly schizophrenic and depressed) who attained spuriously elevated scores on H. The item content of this Ch key was quite puzzling, because although the correction was successful, the items did not seem to refer to anything either hypochondriacal or anti-hypo-

chondriacal. In fact it was difficult to see what psychological homogeneity, if any, they possessed. For a more detailed description of this scale (now no longer in use since the appearance of the modified Hs key) the reader is referred to the original article (42). For present purposes it is merely necessary to state that the great majority of the items on Ch were scored if answered in the statistically rare and obviously "maladjusted" direction and that they apparently measured some non-somatic component of test responses which resulted in spuriously elevated H scores in persons who were not actually hypochondriacal.

Still another scale of the same general sort was derived by Meehl and called N. To briefly repeat what has been said above, this scale differentiated normals showing elevated profiles from clinical abnormals showing no greater profile elevations, and was interpreted as detecting a plus-getting test attitude for which scores on the personality components proper should be corrected. The type of item occurring on the scale N has been discussed above.

Lastly, we recall to mind the Hy-O items which have been described above as reflecting this kind of component, although scored in the opposite direction from N, Ch, and G.

It is of considerable interest to examine the correlations between K and these other variables, derived in their diverse ways. Table 2 presents the correlations between K and the various scales thought to be loaded with the factor in question, based upon scores of 100 individuals ages 26-45 in each of the groups indicated.

Table 2
Correlations of K Scale with Other Variables Thought to be Loaded
with the "K-factor"

	+	G	N	Ch	Hy-O
Normal males	-.64	-.76	-.70	-.67	.81
Normal females	-.62	-.73	-.64	-.63	.78
Male abnormals	-.70	-.75	-.69	-.64	.74
Female abnormals	-.70	-.81	-.72	-.71	.74

Considering the relative unreliability of some of these variables, the above is a very impressive group of intercorrelations. We have two scales (G and +) which were derived wholly by internal item relationships and without regard to criteria of any non-test behavior; a scale (N) which corrects for the self-criticality of certain plus-getters who show deviant profiles; a scale (Ch) which differentiates hypochondriacs from non-hypochondriacal abnormals who have elevated H scores; and a sub-

set of items (Hy-O) which were chosen because they differentiate a clinical group—hysteria. There is, however, a considerable item overlap among these scales, tending to raise these correlations. On the other hand, it will be recalled that the scale K is not actually “pure” for the hypothetical test-taking attitude because it is a composite of the test-taking scale L_6 plus the eight “psychotic” items. This would presumably tend to lower the correlations. Accordingly, we have substituted L_6 for K, removed the item overlap among the scales G, N, Ch, L_6 and Hy-O, and calculated correlations among these reduced keys. Table 3 shows the intercorrelations among these five non-overlapping keys, based upon the responses of 150 unselected normal males between the ages of 26 and 45, rejecting records with $? > 70$ or $F > 80$. All scales were scored so as to render the correlations positive.

Table 3

Intercorrelations of Five Scales Thought to be Loaded with the Test-taking Attitude, No Item Overlap. $N = 150$ Normal Males

	G	Ch	L_6	N
Ch	.82			
L_6	.76	.71		
N	.78	.73	.66	
Hy-O	.70	.63	.70	.59

This correlational matrix has been subjected to a factor analysis, repeated three times in successively approximating the communalities because of the small number of tests. The first factor extracted leaves no residuals larger than .049, and the SD of the residuals is .032, which is less than the SE of .041 attached to the mean r in the matrix. Testing the significance of the residuals by the formula $\chi^2 = \sum (z_0 - z)^2 / (n - 3)$ (12, p. 339) the chi-square on the deviation of observed r 's from those predicted with the first factor loading was not significant ($\chi^2 = 5.101$, 5 d.f., $P > .30$). It appears that one common factor is quite sufficient to account for the intercorrelations of these scales. The factor loadings of the scales G, Ch, L_6 , N, and Hy-O are .927, .868, .847, .818, and .770 respectively. It is interesting to find such a powerful factor running through scales derived by such diverse methods. It is also worth noticing that the largest loading of the K-factor is in the one scale constructed wholly by “internal consistency” methods, whereas the smallest loading is that of the clinical variable Hy-O. If we extract a second factor just to see what it looks like, none of the loadings is over .20 and the meaning of the second factor would be quite uninterpretable on our data. Although we have been thinking in terms of a “K-factor” on

the basis of the apparent community of practical function shown by these various scales, it is reassuring to find that the term "factor" may be used here without doing violence to the more technical meaning of that term as used by factor analysts.

Considering the nature of the items which are involved in scales such as L₆, N, and G, this finding perhaps sheds some light on the relative inadequacy of "neurotic" inventories such as the BI-N when applied to clinically diagnosed neurotics. Here we have a kind of item which, while it does not (in its own right) appear to discriminate normal from abnormal individuals very successfully, does reflect some kind of a test-attitude or self-critical component. Those "neurotic" persons who happen to be characterized by this particular manifestation of self-criticism, such as certain compulsives, will probably be differentiated by such a set of items. On the other hand, other equally "neurotic" persons such as hysterics, who are characterized by the opposite attitude, will not be successfully spotted by the scale. If anything, they should be discriminated backward! Furthermore, the central tendency of abnormals in general is the same as that of normals, and it is quite possible that in developing personality questionnaires set up in the traditional, *a priori* fashion and "refined" by statistical manipulation we are merely setting up sets of items to differentiate among people with respect to various test-attitude continua of little or no psychiatric relevance. It will be recalled that the scale G consisted of items having the heaviest loading with whatever factor (or factors) contribute most to the variance and covariance of the entire 550 items in the MMPI pool. Yet this scale turns out to have little or no clinical value (*except* as a suppressor) and to be the scale most saturated with respect to the test-taking attitude. We feel that psychologists have tended to forget the fact that when one constructs a personality inventory by studying the item-associations, whether by old-fashioned methods of internal consistency or by factor analysis of item correlations, he is merely locating certain covariations in verbal behavior. When a final scale based upon that kind of derivation is presented to the clinician, all that the clinician can be assured of is that *persons who say certain things about themselves also have a tendency to say certain other things about themselves*.

Willoughby's argument (65) that the non-chance covariation of item responses establishes "validity" with respect to *some* underlying, common trait which gives rise to the covariation may be admitted without contradicting what we have just said. That items should exhibit consistency in this covariant sense in spite of not being valid for the traits sought, or in fact even being negatively valid, has been shown by many studies, most particularly those of Landis and his associates (34, 35, 52). The

"underlying disposition" which leads a subject to respond in a certain way to such questions may or may not be identical with the dispositions we recognize as clinical variables, nor with those that might be suggested by the item content. It is quite clear on present evidence that this identification cannot be established by an assumed equivalence between non-test behavior and the verbal report. Hence, as has been repeatedly stressed by the present writers, both *a priori* selection of items and the psychological naming of a statistically homogeneous scale from its item content are fraught with possibilities of error.

An obvious line of investigation which is suggested by these considerations is the systematic study of the relationships which exist among variables such as K, G, and N which are fairly definitely known to be chiefly test-taking variables, and other personality scales which have been developed by variants of the method of internal consistency. Because of the influence of socio-economic or educational level upon the K-factor (see Section VI below) such studies should ideally be carried out upon subjects from the general population. At present, we can only report a few preliminary studies which seem to have some bearing upon this question. All of these studies happen to be concerned with the batteries developed by Guilford and Martin (GAMIN, STDCR, and the Personnel Inventory). We wish to emphasize that the presentation of these scattered data on our part is intended simply to raise some questions concerning the construction of scales by internal consistency methods where factors such as K are probably in operation; the validity of the Guilford-Martin scales must of course be assessed upon other grounds. We wish further to stress that in comparing these tests with MMPI we do not intend to set the latter up as a "criterion," although it does of course have the advantage that each item is known to differentiate certain defined criterion groups which literally define the scales on which the item occurs. It should also be made clear that Guilford, as one of the foremost contributors to the factor analytic approach to personality test construction, has explicitly called attention to the importance of the problem of test-taking attitudes as "factors," when he says,

"We must constantly remember that the response of a subject may not represent exactly what the question implies in its most obvious meaning. Subjects respond to a question as at the moment they think they are, with perhaps a lack of insight in many cases as to their real position on the question. They also respond as they would like themselves to be and as they would like others to think them to be and as they wish the examiner to think them to be. They also respond with some regard to self-consistency among their own answers. Whether these determining factors are sufficiently constant to set up individual differences which are uniform in character and so constitute common factors in themselves is difficult to say. Should any one of them be so pervasive it should introduce an additional vector in the factor analysis" (18, p. 118).

It is our opinion that the data we have presented indicate that the answer to Guilford's question is in the affirmative, and that the inclusion of a few K-type scales in a factor analysis would probably result in a somewhat different interpretation of the other tests and factors than would otherwise be the case.

Wesley (64) has studied the relationships existing between the Guilford-Martin Personnel Inventory of traits O-Ag-Co and the MMPI scales, based upon the test records of 110 presumably normal college women. The three traits measured by the Personnel Inventory are called *objectivity*, *agreeableness*, and *cooperativeness* by their authors. High scores are in the direction of the traits named, and low scores indicate the presence of what is called in composite the "paranoid" personality. Wesley found that the composite Personnel Inventory score correlated only .11 with the MMPI Pa scale which, while still in a preliminary stage, does consist of items which are empirically known to distinguish clearly paranoid groups of persons from people in general. Together with this rather disconcerting finding, she also discovered that the "paranoid" score on the Personnel Inventory correlated .50 and .57 with the MMPI scales Pt and Sc—both of which are relatively weak scales from the standpoint of clinical differentiation but are known to be heavily loaded with the K-factor. The correlations of "objectivity" with Pt and Sc were both $-.62$, which led her to correlate Trait O with the correction scale N, leading to the same figure. None of the other correlations of the Guilford scales with MMPI scales exceeded .45, and the majority of them were under .20. The mean MMPI profile of subjects selected on the basis of having low raw scores on N (the "defensive" end) showed a pattern hardly distinguishable from that of subjects selected for having high scores on Factor O. It is interesting to note in passing that of the seven items of very similar wording which occur on both the Guilford-Martin Inventory and the MMPI Pa scale, five are scored as "paranoid" in the opposite direction on the two scales. For example, to say that most people inwardly dislike putting themselves out to help others, that most people would tell a lie to get ahead, that some people are so bossy and domineering that one feels like doing the opposite of what they tell him to do, are responses scored as paranoid on the Guilford-Martin; whereas it is found empirically that these verbal reactions are actually significantly *less* common among clinically paranoid persons than they are among people generally. This kind of finding suggests that paranoid deviates are characterized by a tendency to give two sorts of responses, one of which is obviously paranoid, the other "obviously" not. But these two sorts of responses are negatively correlated among people

generally, and hence appear scored oppositely on scales developed by internal consistency methods.

It is of course possible to begin the development of scales by internal consistency or item-intercorrelation procedures, and having built a scale by these methods, to apply it to various criterion groups for validation. But it would seem that if the aim is to find items which will optimally perform such a discriminating function, the most direct route to that goal is immediate empirical item selection from the start. It may be agreed that scales developed through item-correlation techniques have more statistical "purity" and hence are in a certain special sense better for what they *do* measure. One's attitude toward this problem is likely to reflect his more fundamental views as to the nature of a so-called "measurement" in personality testing, complete discussion of which would take us beyond the present paper. It seems clear that the results of factor analysis to date have not, whatever their theoretical validity, made possible the construction of single personality items which can be called even approximately "pure." For example, in Guilford's factor analysis of 89 personality items originally chosen (on the basis of suggestions from a previous factor analysis) to sample seclusiveness, thinking introversion and rathymia, after the extraction of nine different factors the majority of the items still showed communalities less than .50. Torrens (61), Wesley (64), and Loth (36) all found that the typical scale intercorrelation among the variables of the Guilford-Martin batteries STDCR, GAMIN, and the Personnel Inventory is actually higher than the typical intercorrelations of scales on MMPI which were developed with almost no consideration for questions of scale purity or freedom from item overlap.

Louis Wesley (personal communication) has suggested that the contrast between the two methods of scale derivation is between *maximal measurement* and *meaningful measurement*. By this is meant that internal consistency methods lead to scales which measure whatever they measure with high consistency, large variance, great discrimination. This is "maximal" measurement. It is suggested that the most important non-test behaviors, which it is the aim of the test to predict, may not be associated with the same variables which lead to the kind of consistency involved. We may, as in the case of the Pa scale, have to sacrifice the desire to have high item intercorrelations in order to score items so as to achieve the more fundamental aim of criterion discrimination. Since scales are so very "impure" at best, there does not seem to be any very cogent reason for sacrificing anything in pursuit of the rather illusory purity involved.

There are multiple determiners which enter into a subject's decision when he answers a personality item. One might say that all but a very

few personality items have an inherently "multiphasic" character, exceptions being such items as "I am a male." Obviously, if there existed or could be invented verbal items which were even approximately pure, the "scales" of such items could be extremely short and in fact the practical value of substituting an inventory for a few brief oral questions would be much in doubt. But the items are not uniquely determined. This simple behavioral fact imposes certain limitations upon the progress of personality measurement, as has been pointed out by many critics. From the common sense point of view, the situation is not very different from what occurs in medical diagnosis or in the psychiatric interview. Almost all of the symptoms or responses which are in evidence are known to arise upon diverse bases. During a psychological interview, a woman may miscall her husband by the name of a former suitor, a phenomenon which is in itself ambiguous; perhaps she has recently seen the man in question, perhaps she has been reading a novel in which that name appears, and perhaps—the psychiatrically significant possibility—she feels somewhat regretful for not having married him instead. Later, we find that she developed a headache on her wedding anniversary, also an ambiguous datum if it stands alone. Again, she is excessively effusive about how happy her married life is, and so on. It is through the hypothesis of marital dissatisfaction that these different behaviors find a common explanation. When we accumulate such single items about her behavior, we are merely piling up the probabilities. It seems a little foolish to locate these behavior particles or their "sum" on a continuum of measurement, except in the most crude ordinal and probability sense. It is further quite likely that important configurational properties are also involved here, so that the significance to be assigned to one of these single facts should be a function of the other facts we know. The traditional scoring procedure of simply counting *how many* responses belonging to a certain class have been made seems to be very crude; fortunately it has been repeatedly found that the various weightings, compositions, and non-linear refinements which the behavioristic logic might suggest do not usually make sufficient practical difference in the ordering and sorting of people to be worth doing. The fact that we find it convenient to treat these behaviors in certain mathematical ways (independent scoring, unit weights, summation, linear transformations, etc.) should not mislead us into supposing that we are doing anything very close to what the physicist does when he cumulates centimeters. From this point of view, methods aimed at either "purity" or "internal consistency" are not easy to justify. At the very best, we have a rather heterogeneous collection of verbal responses which have a rough tendency to covary in strength. It may or may not be true that the most import-

ant (powerful) determiners of this tendency to covary are clinically relevant or personologically significant. For example, disliking one's husband is not the most powerful "factor" in determining the frequency of headaches, among people generally. Nor is it the most potent factor in determining whether one calls him by the wrong name. Furthermore, the tendency to do these two things may not be covariant at all among people in general. None of these reasons, however, would lead us to reject the two facts in trying to evaluate the hypothesis of marital unhappiness.

From both the logical and statistical points of view, the best set of behavior data from which to predict a criterion is the set of data which are among themselves not correlated. This is well known and made use of in the combination of scales into batteries; but for some reason psychologists are uncomfortable if the same reasoning is applied within scales. The statistical considerations are of course quite general, applying as well to items as to scales. It is likely that the insistence upon high internal consistency and "item validity" in the item-test correlation sense springs in part from a feeling that all of the items ought to be "doing the same thing." This certainly sounds like a reasonable demand as it stands, but it requires clarification. As is clear from the factor analysis studies, one simply cannot find any appreciable number of non-identical verbal items which all "do the same thing." Every one of them depends upon many things, and the item as a unit is like the old-fashioned atom—uncuttable and hence permanently impure. Items "do the same thing" when they are so combined in pools that it is very unlikely that the subject will answer many of them in the scored direction unless he is characterized by a certain strength or range of non-test behaviors which in turn depend upon the one (or few) "variables" that are common to the items. It may still (unfortunately) be the case that the heaviest contribution to each item consists of variables other than the ones we are interested in. That this is in fact true is indicated by the typical values of item communalities.

It is this state of affairs which we believe imposes limitations upon the efficiency of such suppressor scales as K. Since we cannot find items which depend upon only clinical abnormality, we try to find items which depend upon abnormality to an appreciable extent even though they unavoidably depend upon other things as well. The suppressor consists of items which unavoidably depend to some slight degree upon clinical abnormality, but to a greater extent upon the objectionable factors in the first set. By cumulating responses to the second set of items, we hope to get an indication of the strength of these other factors, which information is then used to correct for their undesired contribution to a score

attained on the first. The impurity of the suppressor itself, however, sets limits to the efficiency of such a process. Thus, a subject may obtain a high depression score because he is a plus-getter. The strength of his plus-getting tendency is assessed by items such as those of K. However, a sufficiently great degree of depression will yield considerable deviations on K, since the K items themselves are not pure for the plus-getting tendency but are also slightly loaded with clinical abnormality. In such cases K operates against us. It is interesting to note that the K scale, itself a suppressor, also *contains* a suppressor in the form of the eight "psychotic" items—but here also the effort to suppress the unwanted components of the suppressor can only be imperfectly carried out. No refinements of statistical technique enable us to escape the basic psychological fact that our smallest behavior units, the responses made to single items, are inherently of this multiphasic character.

VI. Relation of K to Age, Intelligence, and Socio-Economic Status

In the study of the correction scale N it had been observed that college students (actually, high school graduates tested at the University Counseling Bureau prior to actual matriculation) showed a distinct elevation in the "lie" direction, averaging about one sigma above the general population mean. It was also found that the younger age group (16-25) showed a similar although smaller deviation, which was accounted for by the presence of a considerable number of medical students in that group. Furthermore, college graduates who had been some ten years out of college showed a mean T-score of about 60 on the N-scale. A similar trend is discernible in the case of K. The mean T-score of a group of 84 medical students is at 62, a deviation which is significant at the 1 per cent level. Both male and female pre-college cases average a T of 57 on K. This tendency falls in line with the fact that the mean MMPI curve for several college and pre-college groups, including some obtained elsewhere than at Minnesota, is a curve with a slight but consistent elevation on Hy, in spite of having an Hs below the mean. This indicates, as usual, a tendency to respond in the hysteroid fashion which elevates Hy-subtle enough to more than counteract the tendency to answer the somatic items on Hy in a non-hypochondriacal fashion. We are not prepared on present evidence to give an interpretation of this phenomenon. That it is not primarily a reflection of intelligence differences is suggested by a correlation of only .04 between K and ACE score among the pre-college cases, which, even taking their relative homogeneity into account, should be higher if intellect as such is the reason for the difference. If the factor at work here is not intelligence, nor the mere fact of being in college when tested, two other possibilities are socio-economic status and chronological

age. A group of W.P.A. workers in the young age group 16-25 showed no elevation on K whatsoever, which would favor the socio-economic interpretation. The mean K of a group of 50 normals aged 16-25, excluding college graduates and persons in college, was 13.5 ($T = 52$). These figures would seem to eliminate mere chronological age as the chief basis of differentiation. We are left with socio-economic status as the most plausible remaining variable. What is needed is study of a group of persons in the upper socio-economic group who are not college students and have never been college educated. Unfortunately, we do not have a large enough sample of such persons to enable us to draw conclusions with certainty. The mean raw score on K for a group of 18 normal subjects classified in Groups I and II in the Goodenough classification, who were not, however, college graduates or attending college, was 18.50, which corresponds to a T of 61. In spite of the small N , this difference is great enough so that a t comparing their mean with that of 156 unselected normals from the other economic classes was highly significant ($t = 6.055$, $P < .01$). It seems plausible that the college, pre-college and college-educated elevation is reflecting chiefly a difference in socio-economic status, although further evidence on this topic should be collected. If this is confirmed by subsequent investigation, it will be interesting to speculate upon the possible ways in which membership in the upper classes generates the particular kind of defensiveness involved.

VII. Summary and Conclusions

The general problem of test-taking attitudes in their effect upon scores obtained on structured personality inventories is discussed. The literature on the subject is briefly surveyed, and a discussion given of the various approaches which have been taken in an effort to solve this problem. The final result of many efforts to derive special scales for measuring various attitudes in the taking of the Minnesota Multiphasic Inventory is presented, with some indication of its validity. The relationship of this scale, called K, to other variables is used as a basis for discussing certain general problems in the theory of personality measurement. Conclusions are as follows:

1. The conscious or unconscious tendency of subjects to present a certain picture of themselves in taking a personality inventory has a considerable influence upon their scores.

2. We may distinguish two directions in this test-taking attitude: the tendency to be defensive or to put oneself in a too favorable light, and the opposed tendency to be overly honest and self-critical (plus-getting). The extremes of these tendencies are deliberate, conscious efforts to fake bad or lie good.

3. The defensive tendency appears to be related to the clinical picture of hysteria, whereas plus-getting is related to the picture of psychasthenia.

4. The MMPI scales L and F, while relatively effective in detecting extreme distortion, do not seem to be sufficiently subtle to detect the more common and often unconscious varieties of defensiveness or plus-getting. It has been found convenient to begin interpretation of L in the range of T-scores 55 or 60; whereas F does not clearly establish invalidity even up to T-score 80 (raw score about 16).

5. By contrasting item frequencies of abnormal persons showing normal MMPI profiles and elevated L scores, with the records of unselected normals, an empirical key called K has been derived which is relatively successful in detecting the influence of disturbing test-taking attitudes and can be used to improve the discrimination between normals and abnormals.

6. In studying the intercorrelations among a group of scales derived by various means but all functioning with some effectiveness to detect such attitudes, it was found that one common factor is sufficient to account for all of the intercorrelations. The scale (G) which has the largest factor loading was derived by a method of internal consistency and without recourse to any external criterion. Since K is the scale being used to measure this factor, the factor in question has been called K-factor.

7. On the basis of these findings and study of the relationship of MMPI to certain of the Guilford-Martin scales, it is suggested that perhaps the construction of personality inventories by means of item-correlation and factor analytic methods leads to the development of tests which are excessively loaded with such test-taking attitudes. The procedure of internal consistency in its various forms is called into question as a profitable method for the construction of personality inventories.

Received July 9, 1946.

References

1. Adams, C. R. A new measure of personality. *J. appl. Psychol.*, 1941, 25, 141-151.
2. Allport, G. W. A test for ascendance-submission. *J. abn. Psychol.*, 1928, 23, 118-136.
3. Allport, G. W. *Personality*. New York: Henry Holt and Co., 1937.
4. Allport, G. W. The use of personal documents in psychological science. *Soc. Sci. Res. Council Bull.*, 1942, No. 49.
5. Arnold, D. A. The clinical validity of the Humm-Wadsworth temperament scale in psychiatric diagnosis. Unpublished Ph.D. Thesis, University of Minnesota, 1942.

6. Benton, A. L. The interpretation of questionnaire items in a personality inventory. *Arch. Psychol.*, 1935, No. 190.
7. Bernreuter, R. G. Theory and construction of the personality inventory. *J. soc. Psychol.*, 1933, 4, 387-405.
8. Bernreuter, R. G. Validity of the personality inventory. *Person. J.*, 1933, 11, 383-386.
9. Bernreuter, R. G. The present status of personality trait tests. *Educ. Rec. Supp.*, 1940, 21, 160-171.
10. Bills, Marion. Selection of casualty and life insurance agents. *J. appl. Psychol.*, 1941, 25, 6-10.
11. Bordin, E. S. A theory of vocational interests as dynamic phenomena. *Educ. and Psych. Meas.*, 1943, 3, 49-65.
12. Burt, C. *The factors of the mind*. New York: Macmillan, 1941.
13. Cady, V. M. The estimation of juvenile incorrigibility. *J. Delinqu. Monogr.*, 1923, No. 2.
14. Eisenberg, P. Individual interpretation of psychoneurotic inventory items. *J. gen. Psychol.*, 1941, 25, 19-40.
15. Eisenberg, P., and Wesman, A. Consistency in response and logical interpretation of psychoneurotic inventory items. *J. educ. Psychol.*, 1941, 32, 321-338.
16. Frenkel-Brunswik, E. Mechanisms of self-deception. *J. soc. Psychol.*, 1939, 10, 409-420.
17. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.
18. Guilford, J. P., and Guilford, R. B. Personality factors S, E, and M, and their measurement. *J. Psychol.*, 1936, 2, 109-127.
19. Guilford, J. P., and Guilford, R. B. Personality factors D, R, T, and A. *J. abn. soc. Psychol.*, 1939, 34, 21-36.
20. Guilford, J. P., and Martin, H. G. *An inventory of factors STDCR*. Beverly Hills: Sheridan Supply Co., 1940.
21. Guilford, J. P., and Martin, H. G. *The Guilford-Martin Personnel Inventory*. Beverly Hills: Sheridan Supply Co., 1943.
22. Guilford, J. P., and Martin, H. G. *The Guilford-Martin inventory of factors GAMIN*. Beverly Hills: Sheridan Supply Co., 1943.
23. Hartshorne, H., and May, M. A. *Studies in deceit*. New York: Macmillan, 1928.
24. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule: I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249-254.
25. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule: III. The measurement of symptomatic depression. *J. Psychol.*, 1942, 14, 73-84.
26. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. New York: The Psychological Corporation, 1943.
27. Hendrickson, G. Attitudes and interests of teachers and prospective teachers. Paper given before Section Q, AAAS, Atlantic City, Dec. 27, 1932 (unpublished).
28. Horst, P. The prediction of personal adjustment. *Soc. sci. res. coun. bull.*, 1941, No. 48.
29. Humm, D. G., and Wadsworth, G. W. The Humm-Wadsworth temperament scale. *Amer. J. Psychiat.*, 1935, 92, 163-200.
30. Humm, D. G., Stormont, R. C., and Iorns, M. E. Combination scores for the Humm-Wadsworth temperament scale. *J. Psychol.*, 1939, 7, 227-253.
31. Humm, D. G., and Humm, K. A. Validity of the Humm-Wadsworth temperament scale: with consideration of the effects of subjects' response-bias. *J. Psychol.*, 1944, 18, 55-64.

32. Kelly, E. L., Miles, C. C., and Terman, L. M. Ability to influence one's score on a typical pencil and paper test of personality. *Character and Pers.*, 1936, 4, 206-215.
33. Laird, D. A. Detecting abnormal behavior. *Jour. abn. Psychol.*, 1925, 20, 128-141.
34. Landis, C., and Katz, S. E. The validity of certain questions which purport to measure neurotic tendencies. *J. appl. Psychol.*, 1934, 18, 343-356.
35. Landis, C., Zubin, J., and Katz, S. E. Empirical evaluation of three personality adjustment inventories. *J. educ. Psychol.*, 1935, 26, 321-330.
36. Loth, N. N. Correlation between the Guilford-Martin Inventory of Factors STDCR and the Minnesota Multiphasic Personality Inventory at the college level. Unpublished Master's thesis, Univ. Minn., 1945.
37. Ludolph, M. The Guilford-Martin Inventory of Factors GAMIN and its relation to the Minnesota Multiphasic Personality Inventory. Unpublished paper, Univ. Minn., 1944.
38. MacKinnon, D. W. The structure of personality. In J. McV. Hunt (Ed.), *Personality and the behavior disorders*. New York: Ronald Press, 1944.
39. Maller, J. B. The effect of signing one's name. *Sch. and Soc.*, 1930, 31, 882-884.
40. Maller, J. B. *Character sketches*. New York: Bureau of Publications, Teachers College, Columbia University, 1932.
41. Maller, J. B. Personality tests. In J. McV. Hunt (Ed.), *Personality and the behavior disorders*. New York: Ronald Press, 1944, 170-213.
42. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule: II. A differential study of hypochondriasis. *J. Psychol.*, 1940, 10, 255-268.
43. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule: IV. Psychasthenia. *J. appl. Psychol.*, 1942, 26, 614-624.
44. McKinley, J. C., and Hathaway, S. R. The Minnesota Multiphasic Personality Inventory: V. Hysteria, hypomania, and psychopathic deviate. *J. appl. Psychol.*, 1944, 28, 153-174.
45. McNemar, Q. The mode of operation of suppressant variables. *Amer. J. Psychol.*, 1945, 58, 554-555.
46. Meehl, P. E. The dynamics of structured personality tests. *J. clin. Psychol.*, 1945, 1, 296-303.
47. Meehl, P. E. An investigation of a general normality or control factor in personality testing. *Psychol. Monogr.*, 1945, 59, No. 4.
48. Meehl, P. E. A simple algebraic development of Horst's suppressor variables. *Amer. J. Psychol.*, 1945, 58, 550-554.
49. Metfessel, M. Personality factors in motion picture writing. *J. soc. abn. Psychol.*, 1935, 30, 333-347.
50. Mosier, C. I. A note on item analysis and the criterion of internal consistency. *Psychometrika*, 1936, 1, 275-282.
51. Olson, W. C. The waiver of signature in personal reports. *J. appl. Psychol.*, 1936, 20, 442-450.
52. Page, J., Landis, C., and Katz, S. E. Schizophrenic traits in the functional psychoses and in normal individuals. *Amer. J. Psychiat.*, 1934, 13, 1213-1225.
53. Rosenzweig, S. A suggestion for making verbal personality tests more valid. *Psychol. Rev.*, 1934, 41, 400-401.
54. Rosenzweig, S. A basis for the improvement of personality tests with special reference to the M-F battery. *J. abn. soc. Psychol.*, 1938, 33, 476-488.
55. Ruch, F. L. A technique for detecting attempts to fake performance on a self-inventory type of personality test. In Q. McNemar and M. A. Merrill, *Studies in personality*. New York: McGraw-Hill, pp. 229-234.

56. Spencer, D. Frankness of subjects on personality measures. *J. educ. Psychol.*, 1938, 29, 26-35.
57. Steinmetz, H. C. Measuring ability to fake occupational interest. *J. appl. Psychol.*, 1932, 16, 123-130.
58. Strong, E. K. *Vocational interests of men and women*. Stanford: Stanford University Press, 1943.
59. Symonds, P. M. *Diagnosing personality and conduct*. New York: Appleton-Century, 1932.
60. Thurstone, L. L., and Thurstone, T. G. A neurotic inventory. *J. soc. Psychol.*, 1930, 1, 3-30.
61. Torrens, J. K. An investigation and evaluation of the Guilford Inventory of factors STDCR with special reference to the Minnesota Multiphasic Personality Inventory. Unpublished paper, Univ. Minn., 1944.
62. Vernon, P. E. The attitude of the subject in personality testing. *J. appl. Psychol.*, 1934, 18, 165-177.
63. Washburne, J. N. A test of social adjustment. *J. appl. Psychol.*, 1935, 19, 125-144.
64. Wesley, Elaine. Correlations between the Guilford-Martin Personality Factors O, Ag, Co and the Minnesota Multiphasic Personality Inventory at the college level. Unpublished Master's thesis, Univ. Minn., 1945.
65. Willoughby, R. R. The concept of reliability. *Psychol. Rev.*, 1935, 42, 153-165.
66. Willoughby, R. R., and Morse, M. E. Spontaneous reactions to a personality inventory. *Amer. J. Orthopsychiat.*, 1936, 6, 562-575.
67. Zubin, J. The method of internal consistency for selecting test items. *J. educ. Psychol.*, 1934, 25, 345-356.

Book Reviews

Rogers, Carl R., and Wallen, J. L. *Counseling with returned servicemen*. New York: MacGraw-Hill Book Co., 1946. Pp. 159. \$1.60.

This clear, concise little book is intended to be a manual for the training of counselors, particularly that large group of men and women who, during the war and its after-math, have become counselors by virtue of the needs of others rather than as a result of a long-term plan and training program of their own.

Its scope is not, however, as broad as its authors claim. It is not a manual of counseling, but rather a manual of non-directive counseling. As such it is an admirable primer. It begins with a brief discussion of the nature of non-directive counseling, of aspects of the psychology of adjustment most important to that type of therapy, and of the attitude of the non-directive counselor. This introductory material is followed by a description, illustrated by well-selected excerpts from interviews and case summaries, of the methods and processes of non-directive counseling. Two chapters are devoted to vocational and educational counseling and to marital counseling, not because the authors consider them special types, but because many others do. There is an excellent discussion of the use of the casual contact, demonstrating how well it can lend itself to non-directive counseling, followed by a series of exercises in responding to clients' statements and questions which are brought together to give some preliminary "practice in counseling"—an excellent teaching device. An appendix of "further reading," all of which has to do with non-directive counseling or the attitudes of returned servicemen, and a good index complete the volume.

Rogers' viewpoint is already well known to psychologists, and its general nature as expounded in collaboration with one of his former students in this brief manual needs no comment here. Some points, however, have been made more specific in this manual than in Rogers' previous writings, and should be noted because in some cases they sharpen our insights into the counseling process, and because in others they are, in this reviewer's opinion, misleading unless modified.

Seven assumptions or claims are made by Rogers and Wallen, each of which should at least be recognized and, in time, validated or rejected as false.

First, is the assumption that most people are maladjusted and need psychotherapy (pp. 1, 2, 14, 90, 96, and 104—"usually . . . the state-

ment of a vocational or educational problem really disguises a deeper personal problem"). There are already some investigations of this question, e.g. Bragdon's college study, not referred to by the authors, who implicitly rule out the possibility of anyone being well enough adjusted *before* seeing a counselor for his desire for expert help in making a decision to be genuine. Must we undergo psychotherapy each time we want to consult a banker about financing a car, an architect about building a house, or a school counselor about choosing a college?

Secondly, it is assumed that maladjustments are primarily the result of attitudes (pp. 2, 113), the possibility of situational maladjustments being not even considered. Counseling, as conceived of by Rogers and Wallen, therefore consists of enabling the client to become aware of his attitudes. This is an important *type* of problem to be able to handle, and a *type* of therapy all counselors should be able to use. But it is surprising that these authors are blind to the importance of the environment and to the effectiveness of a changed situation in many cases, for those who dealt with cases of "operational fatigue" in military hospitals saw many "cures" when discharge papers or V-J Day were in sight, which were at least as good as those which non-directive counseling might have achieved. Rogers' USO experience apparently did not bring him in contact with these cases in a way to give him such insights, and Wallen's military service was only in the training stages of the war when neurotic reactions were less common and cures even less so, but neuroses have long been known to be a means of getting out of difficult situations and of getting compensation.

The third assumption is a part of the second, namely that acceptance of responsibility is more to be sought after than *good adjustment*. This is implicit throughout the book. The authors frequently point out the need to leave everything up to the client, including the right to refuse counseling (*reductio ad absurdum*: this is successful counseling, as it is an assumption of responsibility by the client). Ability to operate on one's own is certainly a desired outcome of counseling, but there are conditions in which it is impossible, e.g., psychosis, childhood delinquency. Rogers seems not to recognize that some adjustment directed by the counselor may at times need to precede the assumption of responsibility, just as the development of appropriate attitudes often needs to precede the use of information.

Fourth, is the claim that the viewpoint of this book is new (pp. 5, 23-24). Rogers and Wallen are unfair to Rank, Taft, and other relationship therapists, whose views this reviewer, together with many other students, studied and tried out at this institution ten years ago, when Rogers was doing the same in Rochester prior to writing his own books. To say this

is not to minimize the great service he has rendered in clearly and concisely expounding and illustrating this type of therapy, and in giving the work a research basis.

Fifth, it is claimed that traditional (pre-Rogerian) counseling means *pointing out* the steps the client should take (pp. 5, 6, 91, 95). This is unfair to the best professional counselors of almost any school, whether of psychoanalysis, vocational guidance, or other. Such counseling has in general characterized the relatively inexperienced, untrained, and insecure counselor. Rogers' approach safeguards these from this type of error. His misapprehension is perhaps due to an inadequate understanding of what *interpretation* is; his illustrations of interpretation (e.g., p. 146) are not what this reviewer would classify as such. Interpretation is actually non-directive or client-centered rather than counselor-centered, if well done; its objective is to hasten insight. But Rogers and Wallen do not see this (pp. 89, 142, 146). "Traditional" vocational counseling is also misrepresented as accepting the presented problem as the real one (p. 94); this is too often true of vocational and educational guidance as *practiced* by ill-trained, inexperienced, or over-burdened school or counseling service personnel, but not as practiced by the better vocational counselors nor as advocated in the literature.

Sixth, it is claimed that diagnosis and prescription involve judging in terms of oneself, imposing one's own values (pp. 20, 27). The authors, therefore, deny the wisdom of either. But on pp. 103 to 105 they emphasize the need for diagnosis, avoiding the actual term, of the use the client is trying to make of the relationship. The reviewer questions the claim concerning diagnosis at least, and the authors prescribe the type of adjustment *they* value most: client acceptance of responsibility.

Seventh, and last, is the apparent assumption that the counselor needs no training in the use of diagnostic techniques or in educational, vocational, or other types of information, even though the value of such techniques is admitted for some cases, at certain stages, in the chapter on vocational and educational counseling (recognition which seems to be forgotten at other points of the discussion). Even the readings in this "manual of counseling" include no references to works on testing or occupational information.

The above points are made, not because of a desire to decry the significance of this little volume, but because it has such a great contribution to make that its limitations need to be clearly pointed out. Rogers and Wallen have done an excellent job of explaining and illustrating non-directive counseling, clarifying its objectives, methods, and steps. They have pointed out how one can assist *one* person whose problem consists of *several* persons, by working with him alone. They have shown the

usefulness of the casual contact, and *how* to use it. They have shown the need for adequate diagnosis (to give it its right name) before vocational counseling, and one way of doing it. They have written a good discussion of an old and widely used technique of test interpretation. They have made available some excellent instructional aids. The book should be widely read, assimilated, and in time re-written in the light of a better perspective.

Donald E. Super

*Department of Guidance,
Teachers College, Columbia University*

Beaumont, Henry. *The psychology of personnel*. New York: Longmans, Green and Co., 1945, pp. xiii + 306, \$2.75.

The author intended this book to be "a general introduction to the contributions which psychology has made and should continue to make with ever-increasing success to the solution of the problems of personnel management." The book deals also with many of the non-psychological phases of personnel management.

The real value of the book will be found in the chapters on "Training Employees," "The Workers' Health," "Promoting Safety," "Providing Incentives," and "Occupational Adjustment." To these subjects the author contributes a fresh point of view and supplies excellent examples from industrial practice. The author misses several opportunities, however, to point out psychological applications.

The chapters on "Analyzing Jobs," "Selecting Employees," "Working Conditions," and "Merit Ratings" merely restate basic concepts and do that rather poorly. References to the reliability of tests, for example, are superficial and misleading (pages 67 and 71). In the section entitled "Selection Ratio" (page 80) the concept is never explained satisfactorily. Such inaccuracies combined with a lack of facility in expression detract considerably from the value of the book.

The author maintains that occupational maladjustment may be prevented by the proper selection, guidance and training of workers and by proper labor conditions such as fair standards, conditions and hours of work, skilled supervision and effective incentives. In presenting such a case, the author is at his best.

There is much opinion and little proof presented in the book, but this is a criticism of the personnel field more than a fault of the author.

Psychologists and industrial men interested in personnel management should read this book because, in spite of its faults, it casts some new light on a field that is still rather dark.

Charles C. Gibbons

*The W. E. Upjohn Institute for Community Research,
Kalamazoo, Michigan*

Boring, E. G. (Ed.) *Psychology for the armed services*. Washington: The Infantry Journal, 1945. Pp. xvii and 533. \$3.00.

This book aims to outline with a minimum of technical language, but geared to college level, what the whole body of psychological knowledge holds for the military man. It is offered as a textbook and as a handbook of psychology, not simply for instruction but for individual reading and reference. The level of presentation is distinctly higher than the companion volume *Psychology for the fighting man*, both in presentation of principles and in the development of their applications, yet it retains the readable qualities of the latter.

The volume impresses the reviewer as a most commendable achievement for several reasons and distinctly to the credit of psychology as a science and as a profession.

1. It is a job done in the war emergency through the collaboration of some sixty persons and an editor. It should be a final answer to the accusation that psychologists cannot agree among themselves upon anything but spend their time quarreling over their theories.

2. It shows that there is in the subject matter of psychology a considerable and respectable body of facts which have been distilled out of the research of the last fifty years. This is the foundation upon which the science will grow.

3. It proves that these facts are not merely curiosa of the laboratory but rather that just about everything with which the psychologist has busied himself has practical utility. The teacher can now without apology to his students dust off his olfactometers, his aesthesiometers, his tachistoscopes, his pseudoscopes and his color wheels, for they have earned their right to a place in the applied laboratory. Of all the classical experiments the reviewer failed to find use only of warm spots, cold spots, touch spots, etc. Perhaps it is there and he just happened to overlook it.

The war is over. But it was a total war and as such it brought within the scope of the armed services every civilian activity, no matter how specialized. For this reason, the book is as applicable to peace time activities as to the emergency of war. The six chapters dealing with sensory functions might seem to hold material least useful for peaceful living, but planes will still have to take off, fly through difficult weather and land. Men will have to communicate with each other under difficult conditions of hearing and people will still need to find their way through strange territory afoot and awheel. As for efficient methods of working, of learning and of teaching one can find ready use for all that the book has to tell. The same is true of the accounts of personal adjustment, of

vocational guidance and selection, of leadership and morale, of opinion and the forces that make it. Any civilian will profit from reading these chapters.

Of course, there are things to criticise. Some statements are so compact as to be hard to understand, some errors of fact have crept in here and there, some items have more space devoted to them than they deserve, and shadings of meaning have been sacrificed for the sake of brevity. What seems to the reviewer to be the most notable achievement, namely, the almost complete avoidance of controversial issues in the presentation of material, will irk those specialists of one or another point of view which seems to have been disregarded. Where choice had to be made, as for instance in the chapters on motivation and morale and on sex, the authors did just that.

As the author of a civilian textbook on applied psychology, the reviewer doffs his hat to the editor of a worthy competitor.

Albert T. Poffenberger

*Department of Psychology,
Columbia University*

Nesbitt, Murrough de B. *The road to Avalon*. Capetown, S. Africa: Hadder and Stoughton, 1944. Pp. 226.

Barton, Betsy. *And now to live again*. New York: Appleton-Century Co., 1944. Pp. 150.

"Avalon" is the author's dream of a colony where crippled men and women may learn to use their new limbs and regain both physical and mental balance.

The *Road* to this Avalon is the author's life story. At thirteen he lost both legs. Fourteen operations on his stumps, seven successive pairs of artificial legs, dreadful pain, poverty, illness—these are not the real story. The real story is learning to walk, to swim, to dive, to ride, to sail; the conquest of pain and self-consciousness; the winning back of a normal social life. It is no less an inspiring story that it has happened many times before.

And yet I am not so sure of the effect of this book on others of like handicap. The author admits that he ran away from life because he was afraid of it. And there is a sort of implication that his *Wanderlust*, his inability to settle down to one job or one place for so many years, had in it something admirable, instead of being merely a personal idiosyncrasy only incidentally connected with his handicap. Surely, however, understandable it may be in his case or another's, irresponsibility and restlessness are no virtues.

The author hates pity but I am not sure he does not pity himself.

And he writes to inspire in us both pity and admiration. There is, to my taste, a trifle too much of the Jack Horner about it all,—though sitting in a corner he certainly did not.

In short, it requires superb artistry to write about oneself in such a way as to inspire others. Perhaps only scamps can write really good autobiographies; saints and heroes inevitably sound a little smug.

Betsy Barton's book, likewise, is written out of her own experience and suffering. She does not hesitate to tell of her own life in order to make a point. But she is always turning from her own life outward. The result is a less powerful but more wholesome book.

Not that it is uninteresting. Simply and plainly written, with many human interest stories, she makes the age-old point that the primary problem of the handicapped is his attitudes. The unity of the mind and body is forcefully put, though in old fashioned terms. If we really believe in the unity, we should leave off talk of mind and body and speak of the human being, the person.

This is not just to split hairs about words. As long as we are thinking in terms of a mind and a body, we must admit that there are "crippled" bodies. But when we think in terms of a person as a going concern,—eating, digesting, hoping, breathing, planning for the future, loving,—we see that there are many who are indeed handicapped by loss of limb or sight or hearing but who *as persons* are not crippled at all, but are gloriously alive and whole. Many of us, on the contrary, though our members seem sound, are, with our petty aches and pains, our anxieties and our fears, our jealousies and our hates, and our indigestions, really crippled persons.

Inspirational books are hard to write. They so easily get preachy; and preaching is dreadfully in the Bruce Barton family tradition, but here is one which carried its point. It isn't a bad book for a lot of people with no visible handicap. And men and women who have been injured will find here valuable practical suggestions for learning "to live again."

Horace B. English

Ohio State University

Gann, E. *Reading difficulty and personality*. New York: King's Crown Press, 1945. Pp. 149. \$2.00.

In an introductory discussion of reading disabilities and the causative factors involved, the writer stresses the view that the whole personality is involved in reading behavior and that there is a dynamic relation between the reader and the meanings he derives. One should seek, therefore, evidences of difficulty in the adjustment of the personality in relation to reading. The statement that personality disturbances *invariably* accompany reading difficulties, however, is not strictly accurate.

The hypothesis to be tested: "Dynamic processes in the personality organization which determine its means or types of adaptations are related to, and influential in the reading experience. These processes are associated with or may be responsible for the difficulties or retardations." Personality patterns of retarded, average and superior readers were compared. Personality was measured by the Rorschach Test and an inventory. The author depends upon published statements that their reliability and validity are adequate. Other rating scales and interest inventories were devised but were not evaluated for either reliability or validity.

According to the claims of the author, the Rorschach system shows that retarded readers, in comparison with average and superior readers, (1) are emotionally less well adjusted and less stable, (2) have feelings of insecurity, and (3) are socially less adaptable. The retarded reader is resistant to reading experience, reflects less interest in and occupation with reading, and shows signs of an unfortunate teacher-pupil relationship. The retarded reader is seen as "a functioning personality, organized in ways which would seem detrimental to efficiency in learning, especially with reference to reading." The author considers that the normally adjusted child will learn to read with average success in the ordinary school situation. She suggests that personality difficulties of retarded readers are not due to lack of success in reading but come from other environmental influences. These inhibiting personality forces lead to reading disability and resolution of these forces, therefore, should be a first step in remedial work. As a matter of fact, there is nothing in the data to indicate whether reading disability causes maladjustment or vice versa. The author is drawing her conclusions and suggesting implications from results which *might be revealed* in future research and from the bias of her own view that reading difficulties and disabilities are part of a larger organization, the total personality.

The reviewer is suspicious of the applicability of the statistical test employed to evaluate reliability of differences since it yields statistical reliability for microscopic differences. Thus, in comparing the mean *rating* of retarded and average readers for concentration, the difference between means of 3.68 and 3.59 with sigmas of .46 and .63 is .09. The computed critical ratio is 3.00 which is above the one per cent level of significance. When one considers the customary reliability and validity of *ratings*, the difference of .09 does not seem to have *practical* significance.

The author has made a contribution in giving added emphasis to the factor of adjustment difficulties in reading disability. But unfortunately she has gone far beyond her data in discussing the implications of her findings.

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- New careers in industry.* Amiss and Sherman. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 227. \$2.50.
- Breaking the skilled labor bottleneck.* Eugene J. Bengé. Connecticut: National Foremen's Institute, Inc., 1942. Pp. 98. \$1.00.
- How to make a morale survey.* Eugene J. Bengé. Connecticut: National Foremen's Institute, Inc., 1941. Pp. 102. \$7.50.
- Job evaluation and merit rating.* Eugene J. Bengé. Connecticut: National Foremen's Institute, Inc., 1946. Pp. 107. \$7.50.
- Your problem—can it be solved?* Dwight J. Bradley. New York: The Macmillan Co., 1945. Pp. 213. \$2.00.
- A chart for the rating of foremen.* R. D. Bundy. Connecticut: National Foremen's Institute, Inc., 1945. Pp. 8. \$.50.
- Objective and experimental psychiatry.* D. Ewen Cameron. New York: The Macmillan Co., 1946. Pp. 390. \$4.25.
- Personal adjustment.* Knight Dunlap. New York: McGraw-Hill Book Co., Inc., 1946.
- Statistical analysis.* Allen L. Edwards. New York: Rinehart & Co., Inc., 1946. Pp. 360. \$3.50.
- Guidance practices at work.* Erickson and Happ. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 325. \$3.25.
- Industrial management in transition.* George Filipetti. Chicago: Richard D. Irwin, Inc., 1946. Pp. 311. \$3.75.
- Enrollment increases and changes in the mental level.* F. H. Finch. Stanford University: Stanford University Press, 1946. Pp. 75. \$1.25.
- How to evaluate supervisory jobs.* Albert N. Gillett. Connecticut: National Foremen's Institute, 1945. Pp. 90. \$7.50.
- Guide to guidance.* Volume VIII. M. Eunice Hilton. New York: Syracuse University Press, 1946. Pp. 58. \$1.00.
- The biology of schizophrenia.* Roy G. Hoskins. New York: W. W. Norton & Co., Inc., 1946. Pp. 191. \$2.75.
- People in quandaries.* Wendell Johnson. New York: Harper & Brothers 1946. Pp. 532. \$3.00.
- Counseling techniques in adult education.* Paul E. Klein and Ruth E. Moffitt. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 185. \$2.00.

- How to handle labor grievances.* John A. Lapp. Connecticut: National Foremen's Institute, 1946. Pp. 294. \$4.00.
- People and books.* Henry C. Link and Harry Arthur Hopf. New York: Book Industry Committee, Book Manufacturers' Institute, 1946. Pp. 166. \$10.00.
- Psychiatry for social workers.* Lawson G. Lowrey. New York: Columbia University Press, 1946. Pp. 337. \$3.50.
- Psychology in industry.* Norman R. F. Maier. Boston: Houghton Mifflin Co., 1946. Pp. 463. \$3.00.
- How to select foremen and supervisors.* R. C. Oberdahn. Connecticut: National Foremen's Institute, 1944. Pp. 60. \$2.00.
- An introduction to educational statistics.* Charles W. Odell. New York: Prentice-Hall Inc., 1946. Pp. 270. \$3.50.
- Modern clinical psychology.* T. W. Richards. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 340. \$3.50.
- Manual of advisement and guidance.* Ira D. Scott. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1945. Pp. 233. \$1.25.
- Sex and the social order.* Georgene H. Seward. New York: The McGraw-Hill Book Co., Inc., 1946. Pp. 301. \$3.50.
- Job evaluation and employee rating.* Richard C. Smyth and Matthew J. Murphy. New York: McGraw Hill Book Co., Inc., 1946. Pp. 255. \$2.75.
- The personnel primer.* Charles S. Stevenson. Connecticut: National Foremen's Institute, 1945. Pp. 32. \$.25.
- Child psychology for professional workers.* Florence M. Teagarden. New York: Prentice-Hall, Inc., 1946. Pp. 613. \$3.75.
- Living issues in philosophy.* Harold H. Titus. New York: American Book Co., 1946. Pp. 436. \$3.25.
- An international convention against antisemitism.* Mark Vishniak. New York: Research Institute of the Jewish Labor Committee, 1946. Pp. 135. \$2.50.
- Proceedings of the 1945 annual conference of the Life Office Management Association.* New York: Life Office Management Association, 1945. Pp. 241. \$5.00.
- Ohio State and occupations.* The Occupational Opportunity Service. Columbus: The Ohio State University Press, 1945. Pp. 198. \$1.50.
- Training supervisors in human relations.* Policyholders Service Bureau. New York: Metropolitan Life Insurance Co., 1946. Pp. 53. Gratis.
- Selection of sales personnel and aptitude testing.* The Society for the Advancement of Management. New York: Sutton-Malkames, Inc., 1945. Pp. 137. \$4.00.

Journal of Applied Psychology

EDITED BY: DONALD G. PATERSON, UNIVERSITY OF MINNESOTA

Consulting Editors

AUL S. ACHILLES, *Psychological Corporation*; WALTER V. BINGHAM, *A.G.O., War Department*; AROLD E. BURTT, *Ohio State University*; ARTHUR I. GATES, *T. C. Columbia University*; JOHN G. JENKINS, *University of Maryland*; IRVING LORGE, *T. C. Columbia University*; QUINN MCNEMAR, *Stanford University*; WILLARD C. OLSON, *University of Michigan*; JAMES P. PORTER, *Swarthmore, Pennsylvania*; EDWARD K. STRONG, JR., *Stanford University*; MORRIS S. VITELES, *University of Pennsylvania*; JOSEPH ZUBIN, *N. Y. Psychiatric Institute*.

Table of Contents

<i>Worker Attitudes Toward Scheduling of Industrial Music:</i> W. A. KERR	575
<i>Analysis of Two Point-Rating Job Evaluation Plans:</i> R. C. ROGERS	579
<i>The MacQuarrie Test for Mechanical Ability: I. Selecting Radio Assembly Operators:</i> C. H. GOODMAN	586
<i>Correlation Between Scores on Ortho-Rater Tests and Clinical Tests:</i> J. DAVIS	596
<i>Occupational Differences in the Minnesota Multiphasic Personality Inventory:</i> W. M. VERNIAUD	604
<i>The Effect of an Increasingly Well Defined Criterion on the Prediction of Success at Naval Training School (Tactical Radar):</i> D. B. STUIT AND J. T. WILSON	614
<i>Prediction of Achievement in Typewriting and Stenography in a Liberal Arts College:</i> D. M. BARRETT	624
<i>Readability of Mixed Type Forms:</i> M. A. TINKER AND D. G. PATERSON . . .	631
<i>Recombination of Ideas in Creative Thinking:</i> L. WELCH	638
<i>Questionnaire and Interview in the Neuropsychiatric Screening:</i> D. H. HARRIS	644
<i>Standardization of the Revised Beta Examination to Yield the Wechsler Type of IQ:</i> ROBERT M. LINDNER AND MILTON GURVITZ	649
<i>Book Reviews</i>	659
<i>New Books, Monographs, and Pamphlets</i>	667

Published Bi-monthly by The American Psychological Association, Inc.
Prince and Lemon Sts., Lancaster, Pa., and
1515 Massachusetts Ave., NW, Washington 5, D. C.

Journal of Applied Psychology

Vol. 30, No. 6

December, 1946

Worker Attitudes Toward Scheduling of Industrial Music

Willard A. Kerr

Tulane University

It is possible, though not necessarily true, that the average factory worker is the best authority on whether or not he should have music when he works, how much he should have, how he should have it, and when he should have it. At least, his opinions on these topics are important and should be investigated. Already it has been demonstrated that factory workers want music (2), that music helps certain aspects of morale (1), and that music increases net worker output in monotonous operations (3).

Actual programming of music for factory audiences, with special reference to the time factor, now usually is done in one of the following ways by the plant broadcasting director:

1. *Fatigue Dip Periods.* In some factory operations a temporary decline in output typically appears at about the middle of each half of the work spell. Some plants schedule most or all of their music programs at these periods of believed fatigue and boredom.
2. *Regular Interval Programs.* Many plants set up a regular recorded music broadcast schedule which provides for 15, 20, or 30 minutes out of every hour of the work shift.
3. *Employee Request Programs.* A few plants do not follow a definite time schedule, but play records as they are requested by employees. In one such plant the music, apparently well received, plays almost continuously.

While advocates of the various methods report favorable results, it seems that no attempt has been made to evaluate preferences of workers for alternative methods in the time scheduling of music programs. The average worker's opinion is a fact which must be regarded as important in evaluating the various methods, because the subjective fatigue-boredom curve does not necessarily coincide with the familiar daily average

hourly production curve, and factors other than fatigue and boredom may condition employees' time desires for music.

Using the tear method of response described elsewhere (4), a *Music Timing Ballot* was designed and administered to three groups of factory employees of the RCA Victor Division, Radio Corporation of America. All were accustomed to work to music. These 666 subjects represent a group of 79 females and 138 males engaged in coil winding machine operations, plus a group of 99 females and 32 males engaged in pressing phonograph records in a Camden, New Jersey, factory, and 291 females and 7 males engaged in assembling radio tubes in a Harrison, New Jersey, plant. Twenty failed to indicate sex. Average age of the employees in the miscellaneous group is 30.6, in phonograph records 37.1, and in radio tubes 25.3. The miscellaneous and phonograph record manufacturing group heard a combination of the first two methods of programming mentioned above while the tubes group experienced the third method.

Per cent of employees in each of the three groups giving a response to each question is indicated in Table 1. The responses to "How much music do you want on your work floor?" tend toward bimodality although a majority of respondents, except in the miscellaneous group, indicate a desire for eight hours of music out of an eight-hour work shift. The average worker wants between six and seven hours of music in eight hours of work.

A plurality of workers, if they were to receive three hours of music daily, want it divided into sixteen sessions, but the average worker wants approximately ten sessions.

In response to "When do you want it?" a distinct tendency appears for the two middle hours of each half of the work shift to receive more votes than the first, pre-lunch, post-lunch, or closing hour of the shift. Music is least desired immediately before and immediately after lunch. These subjective reports, probably based on feelings of fatigue and boredom, are particularly significant in view of the known tendency in many factory operations for output to decline temporarily toward the middle of each half of the work spell.

Tetrachoric intercorrelations among the time variables, sex, and age for all 666 subjects are shown in Table 2. Items one (how much) and two (number of sessions) come nearest of any two items to measuring the same thing, that is, a general liking for industrial music and it is not surprising that the correlation between these two items is .66. Apparently morning or afternoon preference for music (Item 3) is not related with liking for music (Items 1 and 2). Older employees tend to care slightly less for industrial music and females seem to want more of it than do males. It is true, however, that the mean age of the males reporting is

Table 1

Preference of 666 Factory Workers for Arrangement and Timing of Broadcast
"Music While you Work"

<i>N</i> =	Per Cent of Employees Giving Each Response to Each of Three Major Timing Questions			
	224 Coil Winding	135 Phono Pressing	307 Tube Assembly	666 Total
<i>1. How much music do you want on your work floor?</i>				
A Never	00.5	00.0	00.6	00.5
B Lunch and rest periods only	01.0	00.8	00.0	00.5
C One hour out of eight	00.0	03.9	00.0	00.8
D Two hours out of eight	05.9	02.3	00.3	02.5
E Three hours out of eight	09.3	03.1	01.0	04.0
F Four hours out of eight	19.5	17.0	06.1	12.5
G Five hours out of eight	03.9	02.3	07.6	05.4
H Six hours out of eight	10.7	02.3	11.1	09.2
I Seven hours out of eight	01.5	03.1	09.6	05.7
J Eight hours out of eight	47.8	65.1	63.69	58.9
Total	100.0	100.0	100.0	100.0
<i>2. How do you want it?</i>				
A All in one session in the first half of shift	01.0	00.0	02.3	01.5
B All in one session in the second half of shift	00.5	00.0	00.3	00.3
C All in two sessions, one in the first half and one in second half of shift	10.2	09.3	07.2	08.5
D All in four sessions—one session of music in every two hours of work	21.8	18.5	15.3	18.0
E All in eight sessions—one session of music in every hour of work	35.4	23.2	30.0	30.6
F All in sixteen sessions—one session of music in every half hour of work	31.1	49.1	44.9	41.1
Total	100.0	100.0	100.0	100.0
<i>3. When do you want it?</i>				
A First hour	09.2	07.0	11.8	10.4
B Second hour	09.8	09.0	15.3	12.7
C Third hour	19.4	12.0	18.1	17.9
D Fourth hour	06.6	16.0	07.3	07.9
E Fifth hour	05.5	06.0	07.1	06.4
F Sixth hour	18.5	11.0	19.5	18.3
G Seventh hour	16.5	24.0	11.4	14.5
H Eighth hour	14.5	15.0	09.5	11.8
Total	100.0	100.0	100.0	100.0

Table 2
Tetrachoric Intercorrelations Among Five Items on the Music Timing Ballot
for 666 Factory Workers

	2	3	5	6
1. How much	.66	.03	.40	-.28
2. How (sessions)		.00	.26	-.27
3. When (afternoon)			-.06	.02
5. Female sex				-.59
6. Age				

significantly higher than that of the females. Also, some older males tended to have jobs involving more supervisory responsibilities. These latter facts must be considered in interpreting the two following partial correlations. Correlation of amount of music desired with sex when age is held constant by technique of partial correlation is .30, and a similar correlation of amount desired with age when sex is held constant is $-.06$. These results indicate that sex (female) more than age is a determinant of how much music a factory worker wants to hear while working; however, it again must be emphasized that sex in itself may be less of a real causal factor than the fact that work performed by the average male subject in this study is of a less monotonous nature than that performed by the average female employee.

Received May 3, 1944.

References

1. Middleton, W. C., Fay, P. J., Kerr, W. A., and Amft, F. The effect of music on feelings of restfulness—tiredness and pleasantness—unpleasantness. *J. of Psychol.*, 1944, **17**, 299-318.
2. Kerr, W. A. Three studies in plant music. *Factory Management and Maintenance*, 1943, **101**, No. 11, 280-286.
3. Kerr, W. A. Experiments on the effects of music on factory production. *AAAP Monogr.*, 1945, **5**, 1-40.
4. Kerr, W. A. Where they like to work; work place preference of 228 electrical workers in terms of music. *J. appl. Psychol.*, 1943, **27**, 438-442.

Analysis of Two Point-Rating Job Evaluation Plans

R. C. Rogers

De Laval Steam Turbine Company, Trenton, New Jersey

The primary aim of any job evaluation system is to provide management with a valid measure of relative job worth upon which to build its wage structure. *From the standpoint of measurement*, the first and most important task of job evaluation is the construction of a battery of discriminative measures which, when properly weighted, will furnish a reliable index of the relative value of all jobs in the population being analyzed.

Underlying most of the existing systems of job evaluation are the assumptions (a) that the job evaluation plan provides measures of "job" characteristics rather than "employee" characteristics, (b) that each of the factors provides a discrete measure of some aspect of job worth and that they are capable of independent evaluation, (c) that each of the factors bears a significant association with the total measure of job worth, (d) that each factor is "weighted" in proportion to its unique contribution to the total evaluation, and (e) that the plan includes all or most of the significant "common denominators" of job worth.¹

A recent factor analysis of the eleven-factor NEMA method of job evaluation (1, 5) by Lawshe and Satter (2) demonstrates that a number of the above assumptions are untenable. Their results indicated that "most of the variance in total point ratings" could be accounted for by one primary factor, "Skill Demands," which was made up of attributes or characteristics possessed by the successful employee. The second factor isolated in their analysis, "Job Characteristics," was made up of physical characteristics of the job itself "with which the employee must contend." Their criterion measure, total points, did not have a significant loading on this factor.

The correlations in their study revealed that certain of the variables, e.g., working conditions and physical demand, were not significantly associated with total points and that a number of the factors did not provide unique measurements.

This paper presents the results of a similar statistical analysis of two

¹ The assumptions of reliability and validity underlying these systems are not treated in this paper and hence are not included in this list.

other job evaluation plans adapted for use in the evaluation of wage (hourly-rated) and salary jobs (6).²

The Job Evaluation Plans

The job evaluation plan for wage (factory) jobs provides for the point-rating of the following *six* factors (numbers in parentheses are the maximum point values possible for each factor): Mentality (100); Skill (400); Responsibility (100); Mental Application (50); Physical Application (50); and Working Conditions (100).

The job evaluation plan for the salary (office) jobs provides for the point-rating of the following *ten* factors: Mentality (150); Training (300); Analytical Ability (300); Initiative (300); Personal Requirements (300); Executive Responsibility (325); Monetary Responsibility (260); Dependability and Accuracy (65); Mental Application (70); and Physical Application (30).

The "total point" rating for each job is obtained by summing the values assigned to the individual factors and adding a constant 400 "base points" to this total. These total ratings are then translated into Job Grades which encompass defined ranges of total point values.

Results

The absolute point-values assigned to each job evaluation factor and the Job Grade for 170 wage (factory) jobs and 295 salary (office) jobs were coded and punched in I.B.M. cards. In addition, the following variables taken from the job descriptions (but not included as such in the job evaluation plans) were coded and punched: for the wage jobs—Learning Time and Educational Requirements; for the salary jobs—Learning Time. The two populations of jobs were treated separately.

Intercorrelations among the variables in the wage job evaluation plan are presented in Table 1; for the salary plan in Table 3. These matrices were further analyzed by means of Thurstone's Centroid factor analysis technique (4). Factor loadings for the variables in the wage plan are given in Table 2; for the salary plan in Table 4. Maximized multiple correlation coefficients (3) were computed from these data in order to determine the best battery of measures in each plan.

Discussion

Throughout the discussion of these results, it is important to bear in mind the following major limitations of such a study: (a) To some extent,

² This study is one of a series conducted in 1944 in connection with a research program aimed at the analysis of existing job evaluation systems in terms of their adequacy as measuring instruments.

the magnitude of the correlations between the factors in each plan and Job Grade (an "internal" criterion) is a function of the *a priori* weights assigned to the factors; (b) The reliability of the assigned point-values is unknown; (c) There is no estimate of the statistical validity of these measures in terms of an *external* criterion of job worth.

Wage Job Evaluation Plan. Considering only the six variables included in the wage plan, it is evident (Table 2) that Factor I accounts for most of the variance in Job Grade. Following Lawshe and Satter's terminology (2), this factor might be named "Skill Demands." The characteristics having high loadings on Factor I, Skill, Mentality, Mental Application and Responsibility, might be taken to represent those characteristics which the employee must bring to the job in order to perform it successfully.

Table 1
Intercorrelations—Wage Jobs
N = 170

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) Job Grade	.88	.96	.87	.87	.01	.15	.89	.46
(2) Responsibility	—	.86	.88	.84	-.17	-.05	.83	.61
(3) Skill		—	.88	.86	-.11	-.00	.86	.50
(4) Mentality			—	.83	-.17	-.15	.80	.53
(5) Mental Applic.				—	-.17	-.04	.82	.55
(6) Physical Applic.					—	.56	-.04	.39
(7) Working Condit.						—	.11	-.20
(8) Learning Time							—	.47
(9) Educational Req.								—

Table 2
Factor Loadings—Wage Jobs

	Unrotated				Rotated			
	I	II	III	<i>h</i> ²	I	II	III	<i>h</i> ²
(1) Job Grade	.85	.37	.16	.89	.93	.16	-.02	.89
(2) Responsibility	.95	.08	.05	.91	.94	-.16	.01	.91
(3) Skill	.94	.23	.11	.95	.98	.00	-.00	.96
(4) Mentality	.94	.17	.05	.92	.95	-.07	-.03	.91
(5) Mental Applic.	.92	.16	.11	.88	.94	-.06	.03	.89
(6) Physical Applic.	-.22	.68	.32	.64	-.02	.78	-.01	.61
(7) Working Condit.	-.21	.69	.38	.66	.00	.82	.03	.67
(8) Learning Time	.79	.28	.44	.90	.88	.24	.27	.91
(9) Educational Req.	.57	.36	.51	.71	.70	.39	.30	.73

Table 3
Intercorrelations—Salary Jobs
N = 295

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) Job Grade	.84	.90	.92	.95	.89	.77	.87	.74	.78	-.05	.86
(2) Mentality	—	.77	.86	.81	.70	.48	.67	.69	.79	-.24	.77
(3) Training		—	.86	.84	.72	.62	.73	.69	.77	-.06	.95
(4) Analytical Abil.			—	.90	.76	.60	.77	.70	.79	-.12	.86
(5) Initiative				—	.87	.69	.82	.67	.75	-.06	.83
(6) Personal Req.					—	.73	.77	.55	.61	-.01	.70
(7) Exec. Resp.						—	.75	.44	.44	.21	.62
(8) Monetary Resp.							—	.63	.60	.03	.73
(9) Depend. and Acc.								—	.71	-.30	.62
(10) Mental Applic.									—	-.32	.73
(11) Physical Applic.										—	-.03
(12) Learning Time											—

Table 4
Unrotated Factor Loadings—Salary Jobs

	I	II	<i>h</i> ²
(1) Job Grade	.98	.15	.99
(2) Mentality	.88	-.24	.83
(3) Training	.92	.08	.85
(4) Analytical Ability	.94	-.04	.88
(5) Initiative	.94	.12	.90
(6) Personal Requirements	.85	.26	.78
(7) Executive Responsibility	.69	.49	.72
(8) Monetary Responsibility	.85	.25	.78
(9) Dependability and Accuracy	.77	-.29	.68
(10) Mental Application	.83	-.34	.81
(11) Physical Application	-.13	.57	.34
(12) Learning Time	.90	.11	.81

Factor II, with high loadings on Working Conditions and Physical Application, was called "Job Characteristics" and reflects those characteristics inherent in the jobs themselves with which the employee must contend. It will be noted that Job Grade does not have a significant loading on Factor II.

None of the measures in this plan has a significant loading on Factor III. And, although Learning Time, and Educational Requirements show low positive loadings, they too have their highest weights on Factors I and II. The existence of this third factor is probably attribut-

able to these two variables neither of which is included in the job evaluation plan as a separate factor.

The shrunken multiple correlation with job grade is .97 with only two of the variables, Skill and Working Conditions included in the battery. Since Skill alone correlates .96 with Job Grade, the increase due to the addition of Working Conditions cannot be considered significant. *Addition of any other variable to this battery actually decreased the magnitude of this correlation.*³ The point-ratings assigned to the other job evaluation factors therefore contribute nothing to the measurement of relative job worth over and above that already assessed by the ratings assigned to Skill. It will also be noted that the arbitrary weights assigned to the factors do not accurately reflect the magnitude of their association with Job Grade.

Examination of the distribution of ratings for Working Conditions and Physical Application indicates that their low correlations with Job Grade may be attributed to the lack of spread in the ratings. Since most of the factory jobs have approximately the same ratings on these factors it would seem advisable that they be "priced" in the job evaluation plan *as a constant* for all jobs, i.e., combined with the standard 400 "base points" already assigned to each job. Since the correlations computed in this study reflect relationships throughout the entire range, there is still the possibility that these factors may be significantly discriminative with respect to certain categories of jobs and hence could not be assigned a constant point-value for all jobs. Further statistical treatment is needed adequately to check this possibility.

Salary Job Evaluation Plan. In this plan also, most of the variance in Job Grade can be accounted for by one primary factor, "Skill Demands" (Table 4), which reflects "employee" characteristics, i.e., Initiative, Analytical Ability, Training, etc.

Factor II, Job Characteristics, composed primarily of Physical Application, again reflects the demands made upon the employee by conditions inherent in the job. Executive responsibility has a low positive loading (.49) on this factor, but its highest loading (.69) is on Factor I, Skill Demands.

The Multiple Correlation with Job Grade is .99 with only three variables, Initiative, Training, and Mental Application contributing to the multiple. The remaining seven measures therefore contribute little to the effectiveness of this evaluation plan. In fact, since Initiative alone correlates .95 with Job Grade, it seems questionable whether Training and Mental Application contribute enough to the final evaluation to justify

³ That is, addition of another variable to the multiple adds more chance error than actual validity and hence decreases the magnitude of the correlation.

including them. This is especially true when we consider the time and cost of arriving at the final ratings.

General Considerations. Although on the whole, these plans have been employed successfully in the establishment of equitable wage structures, and have proved their value from the standpoint of industrial relations, their effectiveness as *measuring instruments* may be seriously questioned. The results of this study, as well as those of Lawshe and Satter, emphasize the fact that many of the principles and techniques of scientific measurement have been neglected in the construction and evaluation of these plans. As a result, many of the elaborate multi-factored systems currently employed contain a number of components which could be dropped from the battery without significantly affecting the accuracy of the final evaluation.

However, it does not seem reasonable to expect that such a complex criterion as relative job value can be reliably measured by means of a single characteristic as the present studies might indicate. Further investigations must be undertaken with a view to developing a reliable battery of discriminative measures in which each of the factors is capable of making a significant and unique contribution to the total evaluation of job worth.

Summary and Conclusions

This paper has presented the results of a statistical analysis of two point-rating job evaluation plans being employed in a metal machining industry for the valuation of wage and salary jobs. An attempt was made to analyze these plans from the standpoint of their effectiveness as measuring instruments. Within the limitations of the study it may be concluded that:

- (1) The present Job Grades in the wage plan could have been determined from the point-ratings assigned to the Skill component alone (with the possible addition of Working Conditions), and those in the salary plan from the ratings assigned to Initiative, Training, and Mental Application.

- (2) In each plan, one primary (centroid) factor, "Skill Demands," accounts for most of the variance in Job Grade. This factor is composed of those characteristics which the successful employee must bring to the job or be capable of developing on the job.

- (3) The second factor in each plan, "Job Characteristics," is composed of those characteristics inherent in the job itself with which the employee must contend. This factor is not significantly associated with Job Grade in either plan.

(4) Many of the variables in each plan are not capable of independent evaluation in their present form, and the *a priori* weights which have been assigned to them do not accurately reflect the magnitude of their association with Job Grade.

Received December 21, 1945.

References

1. Kress, A. L. How to rate jobs and men. *Factory Management*, 1939, **97**, 60-65.
2. Lawshe, C. H., and Satter, G. A. Studies in job evaluation I. Factor analysis of point ratings for hourly-rated jobs in three industrial plants. *J. appl. Psychol.*, 1944, **28**, 189-198.
3. Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. Pp. 245-252. New York: The American Book Co., 1940.
4. Thurstone, L. L. *The vectors of mind*. University of Chicago Press, 1935: "Primary Metal Abilities," Psychometric Monograph No. 1.
5. *Job rating: Definitions of the factors used in rating jobs—hourly rated occupations*. Chicago: Industrial Relations Department, National Electric Manufacturers Association, 1938.
6. *Job evaluation*. Formal plans for determining basic pay differentials. Studies in personnel policy No. 25, National Industrial Conference Board, September 1940.

The MacQuarrie Test for Mechanical Ability:

I. Selecting Radio Assembly Operators

Charles H. Goodman

Radio Corporation of America

This is the first of four articles describing some experiments with the MacQuarrie Mechanical Ability Test ¹ in a radio manufacturing company. The first of these four articles is concerned with the possible use of the MacQuarrie test for selecting radio assembly operators. The second article will describe the findings obtained as a result of a follow-up study of the 329 subjects tested with the MacQuarrie. The third article will set forth the results of a factor analysis of the sub-tests of the MacQuarrie, while the fourth article will consist of a motion analysis of the MacQuarrie's sub-tests.

The experimental work described in these four papers on the MacQuarrie test was part of a psychological research program which was undertaken for the purpose of finding quick selection methods in hiring radio manufacturing workers. The decision to include the MacQuarrie test in this research program was based upon the two features of this test which are highly desirable in an industrial situation; namely, it is a group test, and, secondly, it is a relatively quick measure requiring approximately 30 minutes to administer.

Selective Capacity of the MacQuarrie Test

Subjects. Three hundred and twenty-nine females, hired by the employment office for radio assembly work during the period of November 1943, to March 1944, served as the subjects of this study. No attempt was made to select ² the population. The subjects were simply the first 329 persons hired during the period mentioned. The ages of the subjects ranged from 16 to 64 years with a mean age of 27.3 years and a sigma of 10.2 years. Their ages scatter into an inverse J curve with a modal value of 112 cases or 34 per cent in the first age interval of 15 to

¹ MacQuarrie, T. W. *MacQuarrie test for mechanical ability*. Los Angeles: California Test Bureau.

² Some selection has, of course, taken place through the employment interview. The factory's location in a small rural town where the population is quite homogeneous has also had, in all probability, some selective influence on the population used in this study.

19 years. Only 50 subjects, or 15 per cent of the total group, were more than forty years of age. All subjects were given the MacQuarrie test immediately after they were hired.

The Job. In order, more fully to comprehend this study, it will be helpful to present a brief description of the work performed by an assembly operator in this factory. The job summary taken from the job analysis schedule describes the job as follows: Assembles radio components, such as tube sockets, transformers and capacitors on chassis to form a complete set; assembles terminal boards and other small assemblies using hand tools; mounts subassemblies on chassis and secures them in place using nuts and bolts or soldering iron and rosincore solder; removes insulation from wires using sandpaper or emery cloth, and tins stripped leads; may specialize in one phase of assembly details.

Training. All radio assembly operators are trained for three days in the Vestibule Training School before assignment to assembly lines. During training the new operators are taught how to solder, crimp (the operation of looping wires into terminals or on lugs), and assemble. The basic tools the trainees learn to use are the soldering iron, screw driver, and pliers. This training is designed to give new operators some familiarity with the tools they must use and some practice on the operations they must perform. At the end of two days most operators have acquired enough skill to use their tools and perform the tasks they have been taught.³

Criterion. On the third day of training the new operators are given a manual test to determine how well they have mastered their instruction. The test consists of three different models, A, B, and C, which they must construct. Beginning with model A, they are allowed sufficient time to construct at least two reproductions of the model and after the specified amount of time has elapsed they are told to stop. The same procedure is followed in constructing models B and C. Before starting they are urged to do their best work. Upon completion of the test they are instructed to select one model which they consider their best reproduction of model A, one of model B, and one of model C.

The instructor determines the amount of work done during the test by each operator and uses this for a quantitative score. Qualitative criteria are then applied to each of the models the operators have selected. The following factors are scored: wire dress; length of wire in crimps; number of turns in crimp; excess solder; insufficient solder; rosin or cold joint; loose joint; neatness; and general appearance of the model.

Performance of the operator on this test carries the largest amount of

³ Peak efficiency is not reached by new operators nor is it expected of them until they have worked for several weeks on the production line.

weight in the instructor's final rating of the new operator. Some consideration is also given to attitude and progress during training. The rating system used in the Vestibule Training School is shown in Table 1. These ratings were used as the *criterion*⁴ of success after the letter grades had been converted into numerical values, which are also shown in Table 1.

Table 1
Vestibule Training School Rating System

Letter Grade	Numerical Equivalent	Description of Letter Grade
E	14	Excellent
E—	13	Excellent with some reservations
G+	12	Very good
G	11	Good
G—	10	Below good
A+	9	Better than average
A	8	Average
A—	7	Average with reservations
F+	6	Better than fair
F	5	Fair
F—	4	Below fair standards
P+	3	Just above poor
P	2	Poor
P—	1	Unacceptable

The writer is aware of the very fine grading used in this rating scale, and of the possibility that it was beyond the ability of the training instructor to discriminate so finely. However, no attempt was made to change the rating system for two reasons. First, it had been in operation in this working situation for some time and the training instructor was quite familiar with it, and, secondly, the ratings were heavily dependent upon the objective criteria based on the performance tests which have previously been described.

Study of the distribution of the 329 ratings on this fourteen point scale shows the modal value of 78 cases (23 per cent) to fall at average or its numerical equivalent of 8.0. The calculated mean of the distribution is 7.9, making it appear that there is close correspondence between mean and mode, and that the distribution is fairly normal. However, there are

* Individual production or quality records are not kept for this job, since no one operator assembles a complete radio chassis. Foremen rate the workers after six weeks of production. However, the use of these foreman ratings as a criterion would have involved many raters, while all subjects in this study were rated by the same Vestibule Training instructor.

two other large peaks in this distribution. Fifty-six cases (17 per cent) fall at Good or its numerical equivalent of 11, and 35 cases (11 per cent) fall at Fair or its numerical equivalent of 5. It would appear from this distribution that the instructor tended to rate heavily on the three categories of Fair, Average, and Good.

A regrouping of the letter grades by the writer by combining P—, P, and P+, as the first step, F—, F, and F+, as the second step, A—, A, and A+, as the third step, G—, G, and G+, as the fourth step, and E—, and E as the fifth step, showed a much more normalized distribution of the ratings. It would appear that the training instructor was not able to discriminate as finely as the discriminations called for on the rating scale. Figure 1 shows the distribution of ratings as made by the Vestibule Training Instructor and the distribution of ratings after they had been regrouped by the writer.

Prediction. The adequacy of the MacQuarrie test in selecting assembly operators was determined by calculating Pearson correlations of the subjects' total test score and sub-test scores with the criterion. These correlations are shown in Table 2.

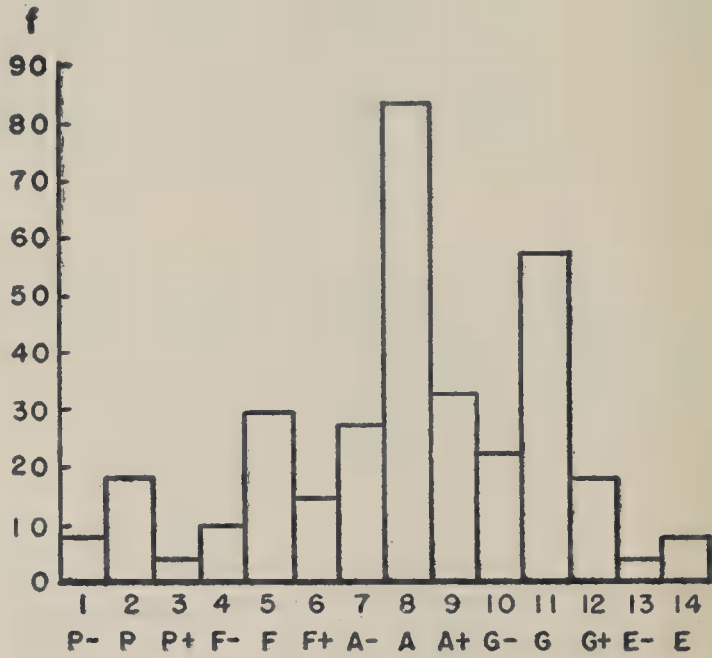
Table 2

Correlations of the MacQuarrie Total Test Score and Sub-test Scores with the Criterion

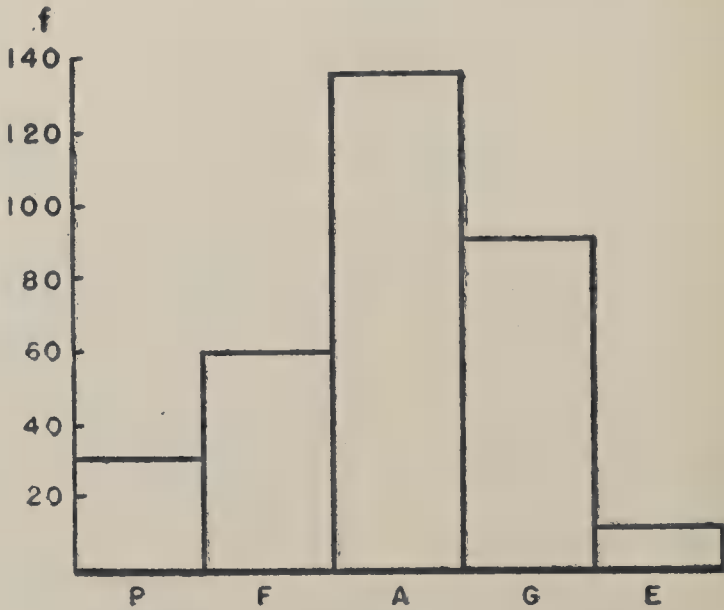
	Total	Part I	Part II	Part III	Part IV	Part V	Part VI	Part VII
<i>r</i>	+.42	+.32	+.18	+.13	+.31	+.35	+.32	+.27
P.E.	.032	.035	.037	.038	.035	.035	.035	.037

Total test score on the MacQuarrie yielded the highest *r* or +.42 with the criterion. Of the sub-tests the highest correlations with the criterion were Location +.35, Tracing +.32, and Copying +.31. To obtain the optimum yield of the MacQuarrie with the criterion, multiple *R*'s were calculated for various combinations of the sub-tests with the criterion. Table 3 shows the multiple *R*'s that were calculated. The criterion *r*'s for the multiple correlations are shown in Table 2, and the inter-correlations are shown in Table 4.

By optimally combining all of the sub-tests, the multiple *R* with the criterion was +.46, which is a small gain over the highest single *r* of +.42. It is interesting to note that a combination of the four sub-tests of Tracing, Copying, Location, and Blocks yielded a multiple *R* of +.44 which is slightly larger than the best single *r*. For quick economical use of the MacQuarrie the three sub-tests of Tracing, Location, and Blocks yielded a multiple *R* of +.41 which is almost as large as the *r* obtained when using the total score of all seven sub-tests of the MacQuarrie.



DISTRIBUTION OF VESTIBULE RATING SCORES



DISTRIBUTION OF VESTIBULE RATING SCORES
REGROUPED

FIG. 1. Distribution of vestibule school efficiency ratings of 329 radio assembly operators.

Table 3

Multiple Correlations of the MacQuarrie Sub-tests with the Criterion

Sub-Tests	<i>R</i>		MacQuarrie Sub-Tests
1, 2, 3, 4, 5, 6, 7	.46	1	Tracing
1, 2, 5, 6, 7	.44	2	Tapping
1, 4, 5, 6	.44	3	Dotting
1, 3, 5, 6	.43	4	Copying
1, 5, 6	.41	5	Location
2, 5, 6, 7	.39	6	Blocks
1, 4, 5, 6, 7	.37	7	Pursuit

Since age was found to correlate $-.22$ with the criterion, a multiple correlation was computed, using the seven sub-tests of the MacQuarrie and age with the criterion. This was done with the thought that since age correlated negatively with the criterion it might contribute towards an increase in the *R*. The resultant *R*, however, showed no increase over the best *R* of $+.46$ which was obtained when using the seven MacQuarrie sub-tests with the criterion. The fact that no increase was obtained when the negative factor of age was added may be explained on the basis of the proportion of "causes" influencing the age factor, the sub-test factors, and the criterion that are common to all.

Apparently the "causes" which produce a negative correlation between age and the criterion are identical with those that produce the negative correlation between age and the sub-tests. Consequently, the addition of age into the multiple correlation does not bring any new correlated "causes" to bear on the relationship.

Age Differences. Question arose as to whether there was any relationship between age and MacQuarrie test scores. To answer this question *r*'s were calculated, and are shown in Table 5.

Table 4

Inter-correlations of the Seven MacQuarrie Sub-Tests

	Tracing	Tapping	Dotting	Copying	Location	Blocks	Pursuit
Tracing		.483	.549	.437	.341	.406	.425
Tapping			.407	.310	.294	.290	.290
Dotting				.340	.430	.320	.360
Copying					.540	.520	.480
Location						.538	.437
Blocks							.459
Pursuit							

Table 5

Correlations of the MacQuarrie Total Test Scores and Sub-test Scores with Age

	Total Score	Part I	Part II	Part III	Part IV	Part V	Part VI	Part VII
<i>r</i>	-.38	-.34	-.23	-.32	-.21	-.23	-.29	-.33
P.E.	.032	.035	.037	.035	.037	.037	.037	.035

All of the correlations are negative and small. However, each r is larger than four times its P.E., indicating that these inverse correlations have some degree of significance greater than zero. It appears then, that as age increases, the MacQuarrie test scores tend in some degree to decrease.

The question of whether a similar relationship existed for age and vestibule training ratings was also answered by correlation. The result showed an r of $-.22$, again an inverse relationship. The P.E. for this r is .037, and similarly the correlation is larger than four times its P.E., which indicates some degree of significance above zero. It would seem then that there is some tendency for the older individuals who enter the Vestibule Training School to receive lower merit ratings.

Superficial inspection of the scattergrams from which these inverse r 's were computed might readily have led one to believe that the relationships were linear. However, plotting of the regression lines raised some question as to the linearity of these relationships. In order to determine how significant these trends were, *etas* were computed.

Table 6 shows the *etas* that were obtained, and the chi-square values which were calculated to test the linearity of the regression, in order to determine whether the curvature might be due to chance deviation from linearity.

Table 6

Eta Coefficients and Chi-Square Test Values for Linearity

Variables	Eta	Chi-Square	Significance
Age and Vestibule Training Rating	.41	46.93	1 per cent level
Age and Total MacQuarrie Score	.64	15.90	5 per cent level
Age and MacQuarrie, Part 1	.42	5.10	Not significant
Age and MacQuarrie, Part 2	.28	6.89	Not significant
Age and MacQuarrie, Part 3	.34	12.72	Not significant
Age and MacQuarrie, Part 4	.40	15.79	Not significant
Age and MacQuarrie, Part 5	.00	10.87	Not significant
Age and MacQuarrie, Part 6	.27	7.40	Not significant
Age and MacQuarrie, Part 7	.38	13.20	Not significant

Table 6 shows that such curvature as may exist in all but two cases may well be due to chance deviation. In the case of age and Vestibule Training rating, the chi-square test shows that the curvature is highly significant being at the 1 per cent level. This shows a highly significant departure from rectilinearity and one would expect similar findings with successive samplings. In the second case where a departure from rectilinearity is shown, that of age and total MacQuarrie test score, the chi-square test shows some significance being at the 5 per cent level.

In view of the significant inverse r 's that were found, it was an expected conclusion that there would be significant differences between the mean test scores of the younger and older subjects. Calculation of the critical ratios of the mean MacQuarrie sub-test scores showed the younger age group to be significantly better than the older age group. These findings confirm the inverse correlation evidence, that the younger subjects of the study do better on the MacQuarrie than the older subjects. This was also found to be true for the ratings received by the younger age group in Vestibule Training School. A critical ratio of 15.00 was obtained showing that the younger age group received significantly higher ratings than the older group.

Efficiency of the MacQuarrie Test. In its most optimum weighting, the maximum correlation of the MacQuarrie Test for predicting future success of assembly radio operators was calculated to be $R = .46$. An evaluation of this R by means of Kelley's coefficient of alination would indicate the effectiveness of the MacQuarrie in predicting individual success of radio assembly operators to be about 12 per cent better than simply hiring without the use of this test.

To the practical plant manager of an industrial concern an increase in efficiency of 12 per cent in hiring might not be too readily accepted or too encouraging in the light of costs for an aptitude testing program.

In order to obtain the maximum efficiency of the MacQuarrie test in this industrial situation for hiring future applicants it was decided that use would be made of the Taylor-Russell ⁵ selection ratio tables. Under ordinary methods of hiring, the plant superintendent judged that approximately 50 per cent of the employees hired were satisfactory. Agreement was reached with the plant superintendent that in future hiring only those individuals who made MacQuarrie test scores that placed them in the top 30 per cent of the distribution of those being considered should be hired.

With an R of .46, the selection ratio set at 30 per cent, and using the estimate that 50 per cent of the employees hired by the old method of

⁵ Taylor, H. C., and Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *J. appl. Psychol.*, 1939, 28, 565-578.

interviewing were satisfactory, the Taylor-Russell Tables indicate that 71 per cent of those selected should be satisfactory when using the MacQuarrie test as a selection device.

The actual testing of this prediction based upon the Taylor-Russell selection ratio was never realized. Shortly after the decision had been made to use the Taylor-Russell selection ratio, the need for employees became so acute, that every applicant seeking work was hired.

A study in retrospect, however, was made of the application of the Taylor-Russell selection ratio, using the same conditions as were to be used for future applicants. These conditions were applied to the group that was originally hired, and will be described in the second of these articles.

Summary and Conclusions

This study has presented data based upon the MacQuarrie Mechanical Ability scores of 329 female radio assembly operators who were hired during the five month period from November 1943 to March 1944 at a radio manufacturing company. The purpose of the investigation was to determine the usefulness of the MacQuarrie Mechanical Ability Test as a selective device for hiring radio assembly operators.

On the basis of the findings of this study, the following conclusions appear to be warranted:

1. The total test score of the MacQuarrie correlates .42 with the criterion.
2. The MacQuarrie sub-test scores correlate with the criterion as follows: Part I, .32; Part II, .18; Part III, .13; Part IV, .31; Part V, .35; Part VI, .32; and Part VII, .27.
3. Part V, the Location test, yields an r with the criterion of .35 which is only .07 less than the total test score yield with the criterion.
4. A multiple R of the seven sub-tests with the criterion shows that when all the sub-tests were used the yield was R .46, which was a slight increase over the total test score of r .42.
5. The use of the four sub-tests of Tracing, Copying, Location and Blocks when optimunly weighted in a multiple correlation yields an R of .44, which is larger than the total test score correlation.
6. The sub-tests of Tracing, Location and Blocks properly weighted yields an R of .41. These three sub-tests produce results almost as large as the seven sub-tests of the MacQuarrie. For economical use of the MacQuarrie in this situation it would appear that these three sub-tests properly weighted would provide results as

good as the total test, plus a saving of time in administration and scoring.

7. The MacQuarrie total test score and sub-test scores were found to correlate negatively with age. These negative correlations are all significant since they are greater than four times their probable errors.
8. It was also found that the relationship between age and total MacQuarrie test score was curvilinear, and that the chi-square test showed this relationship to be significant at the 5 per cent level. It appears, then, that in this particular situation the older subjects tend to make lower scores on the MacQuarrie test.
9. An inverse relationship was also found between age and the criterion of success. The r was found to be $-.22$ and the relationship was definitely curvilinear, the chi-square test showing this curvilinear relationship to be highly significant at the 1 per cent level.
10. The efficiency of the MacQuarrie test for selecting radio assembly operators as interpreted by Kelley's coefficient of alienation would indicate the test to be 12 per cent more effective than the hiring procedure used by the company.

Received December 31, 1945.

Correlation Between Scores on Ortho-Rater Tests and Clinical Tests *

C. Jane Davis

Scientific Bureau, Bausch & Lomb Optical Company

This investigation was conducted to determine the relationship between standard clinical eye tests and the battery of visual skills tests given in the Bausch and Lomb Ortho-Rater. With the increasing use of the Ortho-Rater as a part of testing procedure in personnel and medical departments of industry there has been a growing need for tables permitting conversion of Ortho-Rater test scores to their clinical equivalents.

Procedure

The procedure followed consisted in giving the Ortho-Rater tests to a total of 95 subjects. On each of the individuals a battery of clinical test results were obtained as a matter of routine procedure in an industrial eye clinic.¹ There was no selection of subjects since all individuals reporting for refraction were included. The group consisted of 32 women between the ages of 16 and 57 and 63 men between the ages of 16 and 65. Of these, 46 came into the clinic without glasses and were tested with unaided vision only; 42 either wore or carried glasses and were tested first with unaided vision and then with their present prescription as worn; and 7 wearing bifocals with considerable correction in the distance portion were unable to take the tests without their correction and were tested with glasses only. There were a total of 137 tests run on 95 individuals. In all cases both clinical and Ortho-Rater tests were made in each situation.

Ortho-Rater Tests. Tests on the Ortho-Rater were given in the usual sequence with the addition of monocular tests with the unused eye occluded for right and left eyes at both testing distances. Order of testing was as follows: Distance Tests (Optical Distance of 26 Feet): 1. Vertical

* This article is a "prior publication," the author paying complete costs. The scheduled 80 pages per issue is thereby increased by the corresponding amount, thus the "early publication" of this article is a direct contribution to the subscribers of the *Journal of Applied Psychology* without handicap to those authors whose articles are accepted and printed in their regular turn.

¹ The experiment was carried on in the plant of the Bausch and Lomb Optical Company, Rochester, New York. The facilities of the industrial eye clinic of the plant were used in obtaining clinical results.

Phoria; 2. Lateral Phoria; 3. Acuity Both Eyes; 4. Acuity Right Eye (without occlusion); 5. Acuity Right Eye (monocular with occlusion); 6. Acuity Left Eye (without occlusion); 7. Acuity Left Eye (monocular with occlusion); 8. Depth; and 9. Color. Near Tests (Optical Distance of 13 Inches): 1. Acuity Both Eyes; 2. Acuity Right Eye (without occlusion); 3. Acuity Right Eye (monocular with occlusion); 4. Acuity Left Eye (without occlusion); 5. Acuity Left Eye (monocular with occlusion); 6. Vertical Phoria; and 7. Lateral Phoria. In all acuity tests, subjects were started on the number one target. Following routine procedure,² standard questions were used, explanations and illustrations being supplied where necessary. Every attempt was made to see that the subject understood what was expected of him and that he made his best possible score in each test.

Clinical Tests. Clinical tests were given to all the subjects in this experiment and they included: 1. Acuity Far: Both Eyes, Right Eye and Left Eye; 2. Vertical and Lateral Imbalance (Phoria) Far; 3. Acuity Near: Both Eyes, Right Eye and Left Eye; and 4. Vertical and Lateral Imbalance (Phoria), Near.

Tests were given first without glasses; then if glasses were worn or carried the tests were repeated with prescription as worn. Distance acuity tests were made at 20 feet using a Clason acuity meter. The patient was seated in the examining chair without glasses or wearing his normal prescription and acuity targets were presented following standard clinical procedure. For this series of tests the clinic was asked to find threshold acuity in all cases.

Clinical scores on the Clason have been recorded throughout this paper as ten times the reciprocal of the visual angles. This is identical with Ortho-Rater scoring and results in a value equal to ten times the Snellen decimal.

Determination of Lateral Phoria, Far, employed one of the horizontal lines of letters on the Clason at 20 feet. Two of the testers used a hand Risley prism and the third used a Steven's Phorometer for this measurement. Doubling of the target line is produced by use of vertical prism and the correction in lateral prism required to align the two lines is the phoria measurement. Vertical phoria at a distance of 20 feet was measured in a similar manner using a vertical line of letters on the Clason.

Near point acuity tests were made using a reduced Snellen card which consists of seven rows of letters reduced for Snellen notations at 20 inches. The subject was allowed to hold the test card. This practice was satis-

² *Standard practice in the administration of the Bausch and Lomb occupational vision tests with the Ortho-Rater.* Bausch and Lomb Optical Company, Rochester, New York, 1944.

factory in producing a normal reading posture but resulted in a somewhat varying reading distance since the distance from eye to card was not measured for each individual. Since there was no measurement between a score of 10 (20/20) and a score of 13 (20/15), the scores tended to pile up at 10.

Lateral and vertical phoria at near were measured in the same manner as at distance, using reading card lines as targets.

Results

The statistical procedure adopted for quantifying the relationship between the Ortho-Rater and clinical scores was the determination of Pearson product-moment coefficient of correlation between each of the pairs of tests studied.

Scores on clinical vertical phoria tests approximated a point distribution at orthophoria. Correlations were not run between these tests and the Ortho-Rater tests since prediction cannot be made under these circumstances.

It will be noted in Table 1 that the distance clinical tests correlate with the distance Ortho-Rater tests at a higher level, than do the clinical near tests and the Ortho-Rater near tests. This is probably due to the grossness of the clinical near test for acuity with a pile-up of scores at the score of 10. Right and left eye tests on the Ortho-Rater showed a higher correlation with clinical tests when the occluder was used over the

Table 1
Obtained Coefficients of Correlation
I. Distance Tests

Test	r (Without occlusion on Ortho-Rater)	r (Monocular test on Ortho-Rater)
Acuity Both Eyes	.82	
Acuity Right Eye	.67	.76
Acuity Left Eye	.63	.82
Lateral Phoria	.53	
II. Near Tests		
Test	r (Without occlusion on Ortho-Rater)	r (Monocular test on Ortho-Rater)
Acuity Both Eyes	.71	
Acuity Right Eye	.54	.64
Acuity Left Eye	.67	.70
Lateral Phoria	.64	

eye not being tested than when the test was given without occlusion. In clinical measurement of right and left eye acuity, the tests are necessarily administered monocularly. For this reason the higher correlation demonstrated between clinical scores and monocular scores on the Ortho-Rater is to be expected. In general the distance acuity tests demonstrated correlations of about .80 while near tests gave correlations of about .70.³

Table 2

Regression Equations for Predicting Clinical Acuity from Ortho-Rater Acuity Scores
I. Distance Tests

Test	Equation	S.E. of Estimate
Acuity, Both Eyes	Cl. = .95 O.-R. + .2	1.7
Acuity, Right Eye (without occlusion on Ortho-Rater)	Cl. = .70 O.-R. + 2.4	2.4
Acuity, Right Eye (monocular with occlusion)	Cl. = .98 O.-R. - .2	2.1
Acuity, Left Eye (without occlusion on Ortho-Rater)	Cl. = .58 O.-R. + 3.5	2.4
Acuity, Left Eye (monocular with occlusion)	Cl. = .90 O.-R. + .6	1.8
II. Near Tests		
Acuity, Both Eyes	Cl. = .85 O.-R. + 1.6	1.8
Acuity, Right Eye (without occlusion on Ortho-Rater)	Cl. = .52 O.-R. + 5.0	2.2
Acuity, Right Eye (monocular with occlusion)	Cl. = .80 O.-R. + 2.3	2.0
Acuity, Left Eye (without occlusion on Ortho-Rater)	Cl. = .55 O.-R. + 4.5	2.1
Acuity, Left Eye (monocular with occlusion)	Cl. = .77 O.-R. + 2.1	2.1

From the above correlations and other derived statistics the regression equations in Table 2 for prediction of clinical acuity values from Ortho-Rater scores have been obtained.

Using the regression equations in Table 2, predictive tables have been set up for converting Ortho-Rater scores to their clinical equivalents (Tables 3 and 4). Near and distance tests give somewhat different

³ Test-retest reliabilities run previously on the Ortho-Rater gave coefficients of reliability of between .80 and .90. No reliability values are available for the clinical routine used in the present study; however, the lack of control of the distance on the near test would suggest a low reliability.

values and separate predictive tables are included for each set of tests. In all cases clinical and Ortho-Rater score values are expressed as ten times the reciprocal of the visual angle.

Tables 3 and 4 give values resulting from the interpretation of the regression equations and indicate the mean expected score. An individual attaining a given Ortho-Rater score has a 50% chance of making the predicted clinical score shown in Table 3 or 4, or better. For practical use

Table 3

Table for Predicting Clinical Distance Acuity Scores from Ortho-Rater Distance Acuity Scores

Ortho-Rater Score	Acuity, Both Eyes	Acuity, Right Eye (without occlusion on Ortho-Rater)	Acuity, Right Eye (monocular with occlusion)	Acuity, Left Eye (without occlusion on Ortho-Rater)	Acuity, Left Eye (monocular with occlusion)
1	1.2	3.1	.8	4.1	1.5
2	2.1	3.8	1.8	4.7	2.4
3	3.1	4.5	2.7	5.2	3.3
4	4.0	5.2	3.7	5.8	4.2
5	5.0	5.9	4.7	6.4	5.1
6	5.9	6.6	5.7	7.0	6.0
7	6.9	7.3	6.7	7.6	6.9
8	7.8	8.0	7.6	8.1	7.8
9	8.8	8.7	8.6	8.7	8.7
10	9.7	9.4	9.6	9.3	9.6
11	10.7	10.1	10.6	9.9	10.5
12	11.6	10.8	11.6	10.5	11.4
13	12.6	11.5	12.5	11.0	12.3
14	13.5	12.2	13.5	11.6	13.2
15	14.5	12.9	14.5	12.2	14.1

of the information in referring employees for professional attention a second set of tables (Tables 5 and 6) has been prepared with the predicted value of one S. E. of estimate above the value obtained using the regression equation. In this way, an individual attaining a given Ortho-Rater score has 84 chances in 100 of not exceeding the predicted clinical score. This modified predicted score probably should be used for predicting clinical acuity in cases of employees referred to professional eye men for professional consultation. By using this conservative prediction, there is a marked reduction of the possibility of sending an employee for professional treatment who will achieve higher than the predicted score on the clinical test when given by the professional man.

A similar treatment of lateral phoria scores is not included because (a) the correlations between clinical phoria scores and Ortho-Rater phoria

Table 4

Table for Predicting Clinical Near Acuity Scores from Ortho-Rater Test
Near Acuity Scores

Ortho-Rater Score	Acuity, Both Eyes	Acuity, Right Eye (without occlusion on Ortho-Rater)	Acuity, Right Eye (monocular with occlusion)	Acuity, Left Eye (without occlusion on Ortho-Rater)	Acuity, Left Eye (monocular with occlusion)
1	2.5	5.5	3.1	5.1	2.9
2	3.3	6.0	3.9	5.6	3.6
3	4.2	6.6	4.7	6.2	4.4
4	5.0	7.1	5.5	6.7	5.2
5	5.9	7.6	6.3	7.3	6.0
6	6.7	8.1	7.1	7.8	6.7
7	7.6	8.6	7.9	8.4	7.5
8	8.4	9.2	8.7	8.9	8.3
9	9.3	9.7	9.5	9.5	9.0
10	10.1	10.2	10.3	10.0	9.8
11	11.0	10.7	11.1	10.6	10.6
12	11.8	11.2	11.9	11.1	11.3
13	12.7	11.8	12.7	11.7	12.1
14	13.5	12.3	13.5	12.2	12.9
15	14.4	12.8	14.3	12.8	13.7

Table 5

Table for Modified Prediction of Clinical Distance Acuity Scores from
Ortho-Rater Distance Acuity Scores

Ortho-Rater Score	Acuity, Both Eyes	Acuity, Right Eye (without occlusion on Ortho-Rater)	Acuity, Right Eye (monocular with occlusion)	Acuity, Left Eye (without occlusion on Ortho-Rater)	Acuity, Left Eye (monocular with occlusion)
1	2.9	5.5	2.9	6.5	3.3
2	3.8	6.2	3.9	7.1	4.2
3	4.8	6.9	4.8	7.6	5.1
4	5.7	7.6	5.8	8.2	6.0
5	6.7	8.3	6.8	8.8	6.9
6	7.6	9.0	7.8	9.4	7.8
7	8.6	9.7	8.8	10.0	8.7
8	9.5	10.4	9.7	10.5	9.6
9	10.5	11.1	10.7	11.1	10.5
10	11.4	11.8	11.7	11.7	11.4
11	12.4	12.5	12.7	12.3	12.3
12	13.3	13.2	13.7	12.9	13.2
13	14.3	13.9	14.6	13.4	14.1
14	15.2	14.6	15.6	14.0	15.0
15	16.2	15.3	16.6	14.6	15.9

Table 6

Table for Modified Prediction of Clinical Near Acuity from Ortho-Rater
Near Acuity Scores

Ortho-Rater Score	Acuity, Both Eyes	Acuity, Right Eye (without occlusion on Ortho-Rater)	Acuity, Right Eye (monocular with occlusion)	Acuity, Left Eye (without occlusion on Ortho-Rater)	Acuity, Left Eye (monocular with occlusion)
1	4.3	7.7	5.1	7.2	5.0
2	5.1	8.2	5.9	7.7	5.7
3	6.0	8.8	6.7	8.3	6.5
4	6.8	9.3	7.5	8.8	7.3
5	7.7	9.8	8.3	9.4	8.1
6	8.5	10.3	9.1	9.9	8.8
7	9.4	10.8	9.9	10.5	9.6
8	10.2	11.4	10.7	11.0	10.4
9	11.1	11.9	11.5	11.6	11.1
10	11.9	12.4	12.3	12.1	11.9
11	12.8	12.9	13.1	12.7	12.7
12	13.6	13.4	13.9	13.2	13.4
13	14.5	14.0	14.7	13.8	14.2
14	15.3	14.5	15.5	14.3	15.0
15	16.2	15.0	16.3	14.9	15.8

scores given in Table 1 are too low to justify an attempt at accurate prediction and (b) the low correlations ordinarily found between *any* two phoria tests would seem to preclude the use of such tables for industrial prediction. It should be mentioned, however, that in spite of the low correlations found, when the mean Ortho-Rater score of individuals who demonstrated clinical lateral orthophoria was used as the dividing point on the Ortho-Rater scores, 80% of the subjects gave phoria measurements in the same direction on the clinical and Ortho-Rater lateral phoria tests. This finding would seem to indicate that in a large majority of cases, the Ortho-Rater lateral phoria score does indicate at least the direction of a clinical lateral phoria test finding.

Vertical phoria clinical tests in general do not divide a group as finely as the Ortho-Rater test. The scoring on the Ortho-Rater test has proved to be of value in industrial relations for selecting individuals suited visually for certain occupations. A more gross test appears to be satisfactory for general clinical purposes. Predictive data are not shown here for the same reasons as in the case of lateral phoria.

Summary and Conclusions

A testing program was conducted in an industrial eye clinic in an effort to find the relationship between scores on Ortho-Rater tests and

clinical tests. There were a total of 137 tests (both clinical and Ortho-Rater) run on 95 individuals. Pearson product-moment coefficients of correlations ranged from about .60 to .90. Although these correlations are rather low for individual prediction, a series of predictive tables suitable for the needs of industry has been evolved.

The data justify the following conclusions.

1. Prediction from Ortho-Rater to clinical scores as measured in the present study can be made within reasonable tolerances for all acuity tests.

2. The Ortho-Rater lateral phoria tests indicate the direction of the phorias revealed by clinical tests in 80% of the cases, although the correlations are too low to permit prediction of the amount of the lateral phoria.

3. A measurement of the relation between clinical vertical phoria tests and Ortho-Rater vertical phoria tests could not be made because the clinical vertical phoria test scores approximated a point distribution.

Received July 8, 1946.

Occupational Differences in the Minnesota Multiphasic Personality Inventory *

Willie Maude Verniaud

Department of Tests and Measurements, Houston Public Schools

The Minnesota Multiphasic Personality Inventory (MMPI) was given to 97 women in three contrasting occupations, as an aid in determining whether or not there are occupational differences on this Inventory. It is the purpose of this paper to present findings of the investigation.

The Test

The test consists of 550 statements printed on cards, filed by the subject under guide-cards marked "True," "False" and "Cannot Say." There are three validating indicators, (?), (L), (F), and 9 diagnostic scales: Hypochondriasis (Hs), Depression (D), Hysteria (Hy), Psychopathic Deviate (Pd), Masculinity-Femininity (Mf), Paranoia (Pa), Psychasthenia (Pt), Schizophrenia (Sc), and Hypomania (Ma). The Inventory was developed by a clinical psychologist (Dr. S. R. Hathaway) and a neuropsychiatrist (Dr. J. C. McKinley) as an aid in identifying individuals in need of psychiatric attention. A large proportion of scores made by employed people on such an Inventory would be expected to lie within the "normal" range, since individuals who work for a living are, by definition, "normal" enough to maintain themselves in a paid job. Nevertheless, the selection of a clinical instrument for the investigation was deliberate. Under current psychological theory, the so-called functional disorders may represent extreme forms of personality tendencies present in all of us to varying degrees, and become abnormal only when "out of bounds." If this be true, then an instrument sensitive enough to be of value in identifying extreme deviates may be of value in identifying personality differences among functionally normal individuals in contrasting occupations.

The Subjects

The subjects included 40 clerical workers, 27 department store saleswomen, and 30 optical workers from an industrial plant engaged in making lenses and prisms for naval binoculars.

* Material based on Master's Thesis on file in library of the University of Minnesota, July 1945, prepared under the direction of Professors S. R. Hathaway and D. G. Paterson.

Of the 40 clerical workers, 16 were employed in administrative departments of the City of Houston, 16 in administrative offices of the Houston Independent School District, 8 under various Federal, County or District officials. All worked directly under someone with executive or professional status, had duties including both paper work and personal contacts, and had had two years of experience as a minimum. None of the group were workers whose primary function was to direct others rather than serve a chief directly. All were completely free from responsibility for original decision.

The 27 saleswomen were drawn from three Houston department stores by asking an executive in each store to select ten individuals whom he considered good saleswomen. Three of the 30 so selected did not take the test. Those who did, represent 20 selling departments, as follows: Cosmetics 4, Housewares 2, Boys' Clothing 2, Infants' Wear 2, and 1 each from Lingerie, Men's Furnishings, Basement Ready to Wear, Furs, Corsets, Handbags, Better Dresses, Housedresses, Millinery, Paints, Blouses, Automobile Accessories, Toys, Shoes, China-Glassware.

The 30 optical workers were employed in one of the following production departments: blocking, roughing, emery grinding, polishing, finishing. All whose profiles were used had a minimum of four months' experience in one or more of the above processing operations. All had had some experience in an operation other than the one to which the operator was first assigned. The reason for this last requirement lay in the observed fact that Management provided no trained flying squad or other formal organization to cover emergencies and break bottlenecks, assuming the regular operator group to be sufficiently interchangeable that smooth flow of work could be maintained by reshuffling, also that an experienced operator could change methods as often as technicians modified machinery, materials or techniques. Both assumptions were true of the group as a whole.

Tables 1 and 2 show the distribution of the three occupational groups according to age and education.

Table 1
Ages of Occupational Groups

Age	Clerical Workers	Sales- women	Optical Workers	Total
Under 20	4	1	7	12
20-29	14	1	8	23
30-39	12	11	10	33
40 up	10	13	5	28
n.d.	—	1	—	1
Total	40	27	30	97

Table 2
Education of Occupational Groups

Education	Clerical Workers	Sales- women	Optical Workers	Total
6th grade	0	1	0	1
7th grade	0	0	2	2
H.S. Undergr.	0	4	9	13
H.S. Grad.	23	12	15	50
Coll. Undergr.	11	6	2	19
Coll. Grad.	6	0	2	8
n.d.	—	4	—	4
Total	40	27	30	97

Results

Figure 1 shows graphically the mean T-Score profiles of the clerical workers, saleswomen and optical workers. The heavy horizontal line represents mean score of the normative group on which T-Scale was based.

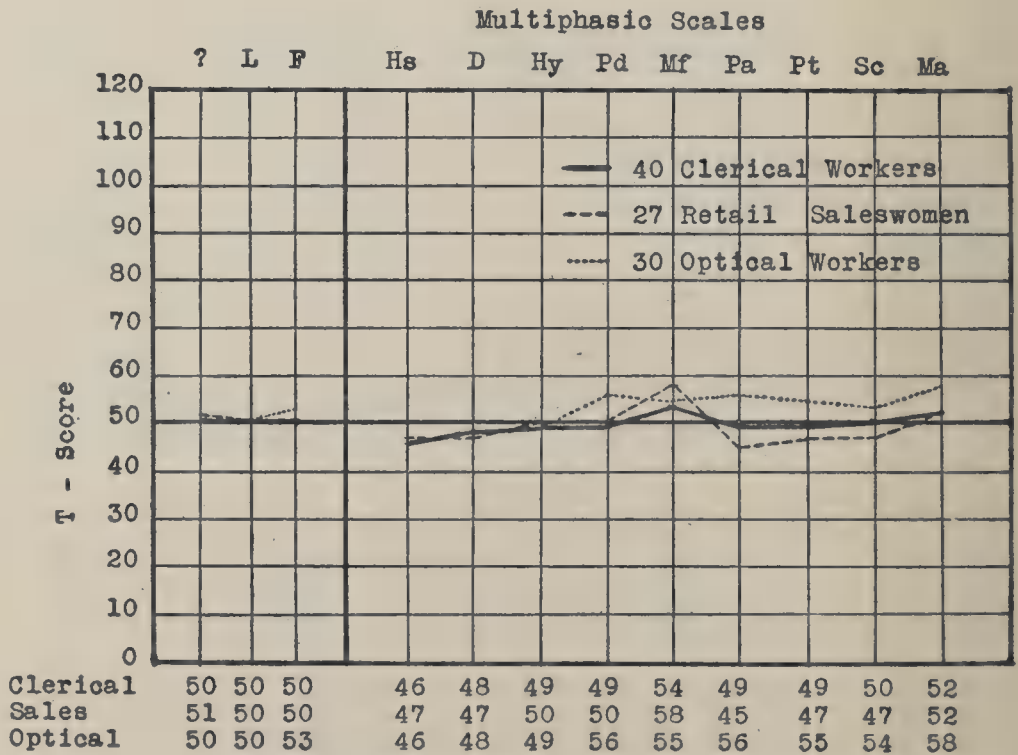


FIG. 1. Mean T-Score profiles on MMPI for the clerical, sales and optical production groups.

The occupational profiles are markedly similar on the first three characteristics of the Scale, the "psychoneurotic triad": Hypochondriasis (Hs), Depression (D), Hysteria (Hy). All fall below or at the norm mean-line. All three occupational means in Masculinity (Mf) and in Hypomania (Ma) fall at some point above the norm mean-line. The clerical worker profile, however, remains reasonably flat throughout and rather closely approximates the norm mean, whereas the composite profile for saleswomen shows a sharp elevation at Masculinity (Mf). As soon as the optical worker profile leaves the "Psychoneurotic triad" (Hs, D, Hy), it mounts to a plateau relative to the mean of the norm group, with Hypomania (Ma) slightly elevated relative to this plateau.

While the profiles show clear group characteristics, they do not yield information as to the probability that the observed differences between occupational and norm means arose solely from errors of random sampling. For this, the reader is directed to Table 3, which shows raw score means, with per cent of occupational group reaching or exceeding norm mean in each characteristic (overlap), standard deviation from mean (S.D.), and ratio of difference to standard error of difference (C.R.). In the case of the following mean differences between occupational groups and the norm group, the null hypothesis can be rejected:

Clerical Workers: Significantly lower mean score in Hypochondriasis;
Saleswomen: Decided differentiation in Masculinity responses;
Optical Workers: Definite differentiation in the direction of Hypomania and Psychasthenia, with statistically significant mean scores in Paranoia and Psychopathic Deviate.

This would mean, for clerical workers, that one would logically expect them as a group to show less tendency toward abnormal concern for bodily functions, less evidence of worry over their health than an unselected sampling of women might show. However, 28% of the group reached or exceeded the norm mean in hypochondriasis, too large a proportion for a low score to have much occupational significance, taken alone. Since no other characteristic is clearly significant, we would expect to find either a tendency to have fewer marked deviations in the direction of abnormality than an unselected group, or so conflicting a set of individual job requirements that the result cancels out in combining the scores in one group profile. In other words, the clerical worker group would appear to be essentially an undifferentiated sampling of the "normal" population.

Three department store saleswomen in 27 (11%), reach or exceed the norm mean in the direction of "femininity." To rephrase, since the T-Scores were reversed for this Scale when applied to females, 24 women

Table 3
Differences between Mean Scores of Norm Group and Occupational Groups on Characteristics of the Minnesota
Multiphasic Personality Inventory

Norm Group *			Clerical Workers N = 40				Saleswomen N = 27				Optical Workers N = 30							
No.	Mean	S.D.	Overlap			C.R.	Overlap			C.R.	Overlap							
			Mean	No.	Per cent		Mean	No.	Per cent		Mean	No.	Per cent					
Hypochondriasis	397	6.86	5.28	4.75	11	28	3.52	3.40	5.44	9	33	4.40	1.06	5.13	12	40	3.63	2.39
Depression	396	19.26	5.18	18.35	21	53	3.73	1.41	17.89	13	48	4.27	1.57	18.50	18	60	4.81	.62
Hysteria	475	18.80	5.06	18.25	18	45	4.54	.72	19.11	14	52	5.57	+	18.13	17	57	4.86	.71
Psychopathic Deviate	397	13.44	4.28	13.33	22	55	3.97	.17	13.70	17	63	3.75	+	16.07	24	80	4.12	+3.29
Masculinity- femininity †	108	36.51	4.83	34.65	14	35	4.23	+2.28	32.52	3	11	3.90	+4.20	34.11	11	37	4.99	+2.32
Paranoia	397	7.98	3.32	7.73	22	55	2.48	.60	6.48	11	41	3.69	2.03	10.00	25	83	2.91	+3.64
Psychasthenia	397	13.08	7.78	11.90	13	33	7.43	.94	11.11	9	33	6.01	1.57	17.03	23	77	6.13	+5.31
Schizophrenia	397	10.73	7.96	10.70	15	38	7.86	—	7.85	7	26	6.03	1.85	14.03	20	67	6.98	+2.43
Hypomania	397	13.65	4.50	15.15	23	58	4.50	+2.00	15.11	18	67	4.40	+1.64	17.47	26	87	4.40	+5.49

* Norm group is group on which T-Scale was based. Data obtained through courtesy of Dr. Hathaway.

** Plus sign (+) in C.R. column indicates that mean difference is in the direction of elevation on the T-Scale.

† Due to the construction of the T-Scale for Masculinity-femininity, a low raw score signifies a tendency in the direction of masculinity in the case of women, with corresponding elevation on the T-Scale.

out of the 27 reach or exceed a T-Score of 50 in the direction of "masculinity." Two of the remainder are at T-Score 49. A solitary score falls definitely at the "feminine" end of the Scale (raw score 44, T-Score 34). This score was investigated, found to represent a pleasant-mannered, middle-aged housewife, one year of college, no work experience outside the home until she entered the present department two years ago, considered an outstanding saleswoman in that location, which is housewares. Despite the fact that these are all department store saleswomen, that is, women who wait on customers who come to them, this is the only definitely "feminine" score found. Would a group of insurance saleswomen, say, or other saleswomen who must seek out their customers, tend to be highly selected in "masculinity"? The answer might help us evaluate what a woman's score in "masculinity" means on MMPI. Unfortunately, there has been little research into the behavioral meaning of a score in this characteristic, when made by a woman. Further, it is the one Scale which was not related to a clinically diagnosed type (for females). To hazard a guess as to interpretation, it might mean a tendency to dominate and direct a situation rather than be dominated by it, a tendency to aggressiveness rather than passivity, and since many of the statements have to do with expressed interests and aversions, a tendency to share "masculine" interests to a greater extent than might be expected in an unselected sampling of women.

Since 26 of 30 optical workers sampled, or 87%, reached or exceeded the norm mean in hypomania, and 77% to 83% of the same individuals also reached or exceeded the norm mean in the other "significant" characteristics: psychasthenia, paranoia and psychopathic deviate, we would be justified in suspecting that average or above-average scores in these characteristics may be related to something in the job, job environment, or job relationships. In terms of the expected meanings of the characteristics, we would expect these workers as a group to be restless, "full of plans," alternating between enthusiasm and over-productivity in energy output and modes of depression, more inclined toward anxieties and compulsive behavior than the average individual, disinclined (or unable) to concentrate for long periods on one task, somewhat oversensitive or suspicious of the good-will of others, somewhat more inclined than the average woman to disregard social mores.

It would not be conservative, however, to draw inferences as to significance in an investigation of this kind on the basis of statistical evidence alone. The next step, therefore, was to study individual data on workers and their jobs. Because of unavoidable limitations, this could not be done for individual saleswomen, and analysis of the other occupational groups consisted of bringing together available material

rather than case study in the clinical meaning. Nevertheless, when the material was assembled, the available clinical evidence was as clear as the statistical in pointing to relationship between type of work, type of worker and characteristic responses on MMPI. Space does not permit the presentation of the bulk of this evidence. The three cases used are illustrative rather than typical. Only five other individuals in the clerical worker group showed as little deviation of score as does the clerical worker used. Both optical workers stand out in their occupational group. The last in particular was sufficiently deviate among her peers that around her small shop legends grew up. Here is also the most deviate profile. The case was selected for its suggestion of close association between rather specific job conditions and the needs of a particular personality.

Illustrative Cases

Clerical Worker No. 40: ?50, L56, F50; Hs54, D49, Hy54, Pd56, Mf47, Pa50, Pt46, Sc39, Ma45

This woman is 38, high school and business school graduate, with some extension work in business administration and in personnel management. She is divorced, shares an apartment with two women friends. She has been working for twenty years, worked throughout the episode of her marriage. Her real love is the department where she began when she finished business school twenty years ago and in which she has worked up to the position of secretary to the director. The director inherited his secretary, who is secure under Civil Service in the job regardless of changing administrations. She appears to have all of the secretarial virtues: an intense, protective loyalty to the man she is serving; disregard of clock hours; discrimination as to what should go to the director and what should be rerouted through someone else; subordination of personalities while on the job; finesse in handling visitors, applicants for favors, complainants, taxpayers, division chiefs and others clamoring for immediate attention from the director. Added to this, she knows the machinery of the department as she knows the palm of her hand.

The most outstanding thing about this worker's attitude is the impression one gets that she has "arrived." There appear to be no unattained horizons, no yearnings for something not yet found. There appears to be also a marked lack of haziness about the various demarcations of life. Things and people belong in the place assigned. Senior clerks, juniors, division chiefs, and so on are accepted with no inquiry into what does not concern her, such as questions of varying qualifications, competence, etc. So long as the person occupies the assigned place, she is as loyal to each, respecting those above and below, as she is to the man appointed as director of "her" department.

Optical Worker No. 5: ?50, L53, F50; Hs41, D42, Hy52, Pd67, Mf63, Pa67, Pt60, Sc58, Ma68

This woman is 34, high school and business school graduate. She is married, with a daughter of 12. She started out as a clerical worker, worked up to the position of secretary to the editor of a local newspaper, advanced from this to newspaper reporter. However, she was put to reporting women's

affairs, had no chance to do general reporting as the men did, and after two years left to work for a clothing manufacturer. She started out as his book-keeper, but also went out into the plant to do cutting. She disliked the book-keeping, "loved" the cutting, and for this reason thought she might like this new occupation for women, when a local university offered vocational courses in optics. She came in 17 months ago when "the whole thing was experimental; so small that one girl did what a department does now." She was "fascinated" by it; and has remained so, has never done anything that gives her such a "kick."

She is now supervisor of the polishers. By the shift foreman's report, she has transformed the polishing room from one of their major headaches to a "bangup" job. She is exacting to work for, unpredictable as a person. She likes to do "screwball" things. Once she emerged from the polishing room with the red polishing compound smeared on hands and arms to the elbows, waved them in the investigator's face, saying "Blood!" During the holidays, when you heard a jingle, you knew that the supervisor of polishers was going through. She had tied bells to her shoes. She is both liked and disliked enthusiastically, but not overlooked. "Finding out it takes a screwball to do this?" she hailed the investigator from across a room. "We oldtimers could have told you, you have to be a little crazy not to go crazy here."

Optical Worker No. 15: ?50, L50, F70; Hs65, D46, Hy56, Pd73, Mf66, Pa65, Pt65, Sc67, Ma81

A second profile was obtained on this 19-year-old girl, as she is the best processor of optical glass in the roughing department, not excepting the roughing supervisor, who admits it. On the second administration, she sat beside the investigator. The cards were read to her, filed as she directed, with essentially the same results as the first time. Despite the high F-score, this is therefore believed to be a valid profile in the sense that it represents a filing of the cards as this subject wished them filed.

She is a high school graduate, finished the college entrance sequence with a B plus average by her statement (probably correct, as the keenness of her observations and comments indicated good scholastic aptitude as definitely as her performance indicated motor aptitude and space-perception equipment). She is unmarried, the daughter of a glass-cutter and glazier, who had his own business and taught her his trade. She was able to handle any operation in roughing, "pick up" any new mechanism and any alteration in procedure, and could teach it to others. The roughing processes, that is, the various operations in the grinding of lenses and prisms, were the processes she limited herself to, but she would be "all over" that department, lending a hand because her quickness would leave her temporarily idle until stations feeding hers could catch up, and was also the first to be reshuffled when absenteeism or other emergency created bottlenecks—until she was fixed to one station by a delicate, complicated new machine, the one machine that "free" operators may not approach lest the operator be diverted. She mastered it. One picked operator at a time was then assigned her to teach, and she was consulted on whether the operator "has what it takes." She liked the honor but disliked the "loneliness" of this station, and when her pleas to be released from this machine went unheeded, she did not remain long thereafter.

"Vivid" and "colorful" are words that come to mind, recalling the impact of this young person on others. Everyone knew her, in and out of her department. Returning to town at midnight, other young people would wait for a bus that she would be on. She usually sang aloud. Yet occasionally, she would have to be roused to her expected role: "Why are you so quiet tonight?" "Can't sing all the time." But presently, she did, the others joining. She

often led the whistling and wolf-calls that greeted embarrassed males who boarded this bus loaded with women.

On last accidental contact, on a city bus, this subject reported that she had passed physical and mental tests and was accepted as a cadet nurse. One wonders what happened after that.

Summary and Conclusions

Samplings of clerical workers, department store saleswomen and women optical workers in a newly opened local industry show characteristic differences in responses on the Minnesota Multiphasic Personality Inventory. These differences are strong enough to indicate that occupational differences in personality, although slight, may be measurable and significant.

The clerical workers approach most closely to a normal sampling, the only clear differentiation being a group tendency toward lower scores in hypochondriasis. The saleswomen are strongly differentiated from the normal sampling in a tendency toward responses designated as "masculine" on this Inventory. The sampling of industrial women deviates from the normal sampling in several respects, a tendency toward hypomania being particularly marked.

These findings must be interpreted in the light of the particular occupational settings. They may not be valid for workers doing similar work, but under very different job conditions. One conclusion can be drawn from this investigation: There are group differences in the personality of successful workers corresponding to gross differences in job requirements, and some of these differences may be identified by responses on the Minnesota Multiphasic Personality Inventory.

Received December 6, 1945.

References

1. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249-254.
2. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule (Minnesota): II. A differential study of hypochondriasis. *J. Psychol.*, 1940, 10, 225-268.
3. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule (Minnesota): III. The measurement of symptomatic depression. *J. Psychol.*, 1942, 14, 73-84.
4. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule (Minnesota): IV. Psychasthenia. *J. Psychol.*, 1942, 26, 614-624.
5. McKinley, J. C., and Hathaway, S. R. The Minnesota multiphasic personality inventory: V. Hysteria, hypomania and psychopathic deviate. *J. appl. Psychol.*, 1944, 28, 153-174.
6. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota multiphasic personality inventory*. New York:

7. Hathaway, S. R., The personality inventory as an aid in the diagnosis of psychopathic inferiors. *J. consult. Psychol.*, 1939, **3**, 112-117.
8. Leverenz, Major C. W. Minnesota multiphasic personality inventory: An evaluation of its usefulness in the psychiatric service of a station hospital. *War Med.*, 1943, **4**, 618-629.
9. McKinley, J. C., and Hathaway, S. R. The identification and measurement of the psychoneurosis in medical practice: The Minnesota multiphasic personality inventory. *J. Am. Med. Assn.*, 1943, **122**, 161-167.
10. Harmon, Lindsey R., and Wiener, Daniel N. Use of the Minnesota multiphasic personality inventory in vocational adjustment. *J. appl.-Psychol.*, 1945, **29**, 132-141.
11. Schiele, B. C., Baker, A. B., and Hathaway, S. R. The Minnesota multiphasic personality inventory. *Journal-Lancet*, 1943, **63**, 292-297.
12. Schmidt, Lt. Col. Hermann O., A.O.D. Test profiles as a diagnostic aid in the Minnesota multiphasic personality inventory. *J. appl. Psychol.*, 1945, **29**, 115-131.

The Effect of an Increasingly Well Defined Criterion on the Prediction of Success at Naval Training School (Tactical Radar) *

Dewey B. Stuit, Lt. Comdr., USNR, and
John T. Wilson, Lt. Comdr., USNR

*Bureau of Naval Personnel and Headquarters, Commander in Chief,
United States Fleet*

This study was undertaken to investigate the validity of the techniques used to select officers for tactical radar training, and to determine the suitability of the criteria of success employed at Naval Training School (Tactical Radar).¹

Research on the validity of selection requirements for Naval Training School (Tactical Radar) was preceded by a National Defense Research Committee exploratory study of techniques employed in the selection of officer personnel for a similar school at St. Simons Island, Georgia. In this preliminary study it was found that the best students were those who made high scores in the Officer Qualification Test, Relative Movement Test, Polar-Grid Coordinate Test, and low scores in the Personal Inventory (Enlisted Form). In addition, it was found that in general the best students were those who had had either administrative or managerial experience and who were judged by trained interviewers to be quick and accurate thinkers. On the basis of these results, selection requirements were tentatively established for Naval Training School

* The opinions expressed in this article are those of the authors and are not to be construed as reflecting the policies or opinions of the Navy Department.

¹ Training in the *tactical* application of radar may be contrasted to training in the *technical* phases of radar, in that the NTSch (Tactical Radar) graduate is assigned to a billet aboard ship in Combat Information Center (CIC). The functions of CIC are *operational* or *tactical* in nature; they are specifically, to keep the Commanding Officer and other higher echelons of command aboard informed of the location, identity, and movement of friendly and/or enemy forces within the area, to control aircraft and small craft in the area, to aid in navigation and to indicate targets. The CIC is manned by a "team" composed of a "CIC Officer" who is in charge, varying numbers of "CIC Watch Officers" qualified to control aircraft and to supervise plotting facilities, and a larger number of enlisted personnel who act as radar operators, telephone talkers, plotters, and communications yeomen. *Technical* radar officers on the other hand are assigned to the relatively more individualized duty of maintenance of electronic equipment, particularly radar. Tactical radar training consists of eight weeks basic training in the tactical employment of radar and in the other related functions of CIC.

(Tactical Radar) the main features of which are: (a) a minimum Navy Standard Score of 50 in the Officer Qualification Test; (b) high scores in the Tactical Radar Aptitude Test² and Relative Movement Test;³ (c) evidence of piloting ability as indicated by preliminary estimate of work in previous navigation courses; (d) maturity of behavior and judgment with a preferred age range from 22-33; (e) educational and occupational backgrounds indicating a high degree of verbal ability, (f) informed volunteer status; (g) physically qualified for sea duty.

Procedure

Selection Tests Used. The selection tests used in this study were as follows: (a) Officer Qualification Test, (b) Officer Classification Test, (c) Tactical Radar Aptitude Test, (d) Relative Movement Test, and (e) CIC Aptitude Test, Form 2.

The Officer Qualification Test was originally designed for use by Offices of Naval Officer Procurement in determining the qualifications of candidates seeking commissions. The test consists of 100 items; 50 Verbal, 20 Mathematical and 30 Mechanical Comprehension. The score is the number right in the total test.

The Tactical Radar Aptitude Test consists of three parts: Polar-Grid Coordinate, Ratio Estimation and Coordinate Reading. The first part measures the examinee's ability to translate the reading of a point on a polar coordinate to a grid coordinate chart. Part two is a test of the ability to estimate relative lengths of lines presented in pairs. Part three measures ability to estimate the direction and range of targets on a polar coordinate chart. There are 50 items in part one, 90 in part two and 90 in part three. Standard scores are computed for the separate tests as well as the total test.

The Relative Movement Test consists of 30 items designed to measure the ability to visualize the relative movement of ships, involving the determination of direction, distance, or speed of ships. Basically the test appears to be a measure of spatial relations ability, but presents problems in a "navigational" setting.

The CIC Aptitude Test, Form 2, consists of three parts: Polar-Grid Coordinate, 45 items, Scale Reading, 60 items, and Relative Movement, 45 items. The Polar-Grid Coordinate and Relative Movement Tests are revisions of the tests of the same name described above. The Scale

² The Polar-Grid Coordinate Test, Ratio Estimation Test and Coordinate Reading Test comprising the Tactical Radar Aptitude Test were originally developed by NDRC Project NS-146.

³ The Relative Movement Test was originally developed by the University of California Division of War Research, Section 6.1, NDRC.

Reading Test, originally developed by NDRC Project NS-146, measures the ability to read scales of various kinds with speed and accuracy. Scores are computed for the three parts as well as for the test as a whole.

The Officer Classification Test was designed for use at reserve midshipmen and indoctrination schools in classifying officers and officer candidates for advanced school training or billet assignments. The four parts of the test are as follows: Verbal, 60 items; Mechanical, 90 items; Mathematical, 45 items; and Spatial, 60 items. Scores are reported only for the parts of the test.

Administration of Selection Tests. The Officer Qualification Test, Tactical Radar Aptitude Test, and Relative Movement Test were administered to all students enrolled in the first two classes at the school. Students reporting from indoctrination or reserve midshipmen schools were tested before their arrival at NTSch (Tactical Radar) and were selected for training upon the basis of their performance in the tests and an appraisal of their personal qualities by interviewing officers. Officers reporting from other shore establishments and fleet commands were tested after reporting to NTSch (Tactical Radar).

The Officer Classification Test was first administered to the third class. During the summer of 1944 the routine administration of this test at indoctrination and reserve midshipmen schools was instituted. Reports of scores made by officers subsequently recommended for tactical radar training were sent to the Bureau of Naval Personnel.

On the basis of preliminary results with the Tactical Radar Test Battery, a revised selection test, the CIC Aptitude Test, Form 2, was constructed. This test was administered at NTSch (Tactical Radar) to approximately one-half of the officers enrolled in the seventh class and to all members of the eighth and ninth classes. The members of the seventh class took the test after they had completed their tactical radar training. Class 8 had completed half of its course of training, and class 9 was in the first week of training when the test was administered.

Criterion Measures Used. The final course grade for the first three graduating classes consisted of a weighted average of the "theory" and "practical" grades assigned each student. The "theory" grade consisted of the arithmetical average of all grades received in weekly quizzes. The "practical" grade was based upon ratings of the student's performance during simulated battle problems, in the school's CICs (Combat Information Centers). The ratings, using a scale of 1 (high) to 5 (low), were made on the following traits: leadership, team work, judgment, mental agility, surface target plotting, air target plotting and speech. The "practical" rating, expressed in terms of the Navy grading system (from zero to 4.0 with 2.5 equaling a minimum passing grade), comprised

two-thirds of the final grade, which was also expressed in the Navy grading system.

In the case of classes 5 and 6, the final grade consisted of the arithmetical average of the "theory," "practical" grades and the added factor of a comprehensive final achievement examination grade. The "theory" and "practical" grades were computed in the same manner as for classes 1, 2 and 3. The final achievement examination grade consisted of the student's score in the CIC Final Achievement Examination, Experimental Form 1. This test consisted of 240 multiple choice items and was administered experimentally to these two classes (5 and 6).

The final grade for classes 7, 8 and 9, consisted of the arithmetical average of the "theory" grade based upon the first month's course grades, a "practical" grade based upon grades made during the second month of the course and performance in a comprehensive "practical examination" and the grade in the CIC Final Achievement Examination, Form 2. The latter test consisted of 240 items and is comparable to the CIC Final Achievement Examination, Form 1, administered to classes 5 and 6.

Method of Analyzing Data. The predictive value of each of the selection tests was determined by correlating test scores with the various criteria of success, using the Pearson product-moment coefficient of correlation as the index of relationship.

Results

Correlations of Aptitude Tests with Criteria of Success

Tactical Radar Aptitude Test Battery. The coefficients of correlation showing the relationship between the tests comprising the Tactical Radar Aptitude Test Battery and the criteria of success are shown in Table 1.

These correlations show that the original Tactical Radar Aptitude Test Battery was not very predictive of success at NTSch (Tactical Radar). In general, the predictive measures correlated higher in the case of classes 5 and 6 than they did in the case of classes 1 and 2. It seems probable that this increase in correlation should be attributed to the changed criteria of success. The results with classes 7, 8 and 9 indicate, however, that the substitution of the second month grade for the "practical" grade did not result in any significant change in the correlations, with the exception of the relationship between the Officer Qualification Test and the achievement test and final grade. This is probably evidence of the fact that the Tactical Radar Aptitude Test and the original Relative Movement Test were in themselves not highly effective predictors of success.

Table 1

Correlation of Measures Comprising Original Tactical Radar Aptitude Test Battery with Criteria of Success at NTSch (Tactical Radar)

The Correlation Coefficients Show by Classes (*) Relationship Between Tests of the Tactical Radar Aptitude Test Battery and Various Criteria of Success. The Number of Cases (**) in Each Class is Also Indicated										
Tactical Radar Aptitude Test Battery	Theory Grade		Practical Grade		Final Grade		Achiev. Test	Final Grade	Achiev. Test	Final Grade
	* 1	2	1	2	1	2	5, 6	5, 6	7, 8, 9	7, 8, 9
	** 97	110	97	110	97	110	103	103	110***	110***
1. Officer Qualif.	.33	.17	.01	.07	.06	.10	.32	.31	.44	.41
2. Polar-Grid Coord.	.19	.03	.29	.20	.31	.19	.27	.19	.26	.24
3. Ratio Estimation	.07	-.19	.26	-.02	.26	-.06	.02	-.08	.17	.21
4. Coord. Reading	-.06	-.14	.18	.13	.11	.08	.34	.28	.07	.02
5. Total (2+3+4)	.10	-.14	.31	.12	.28	.05	.22	.14	.10	.17
6. Relative Movement	.33	.05	.20	-.04	.23	-.03	.25	.22	.23	.18

*** Approximate number of cases varied from 105 to 115 for the six tests.

Officer Classification Test. The scores in the four tests comprising the Officer Classification Test correlated very low with the criteria of success employed with class 3 (see Table 2). Again, results shown in this table demonstrate the effect upon the obtained correlations of a change in the criterion measures. For class 3, only the relationship between the verbal test and the final grade is above the 5 per cent level of significance. For classes 7, 8 and 9 the correlations for both the verbal and mathematical parts are well above the 1 per cent level of significance. It is also to be noted that the correlations with the final achievement examination are somewhat higher than the correlations with final grades. The most striking fact, however, is the high relationship between the mathematical part of the Officer Classification Test and both the achievement test and the final grade. Evidently this factor was much more closely associated with success in later classes than it was in the early classes.

Since the correlations for class 3 are based upon a relatively unrestricted population, and the correlations for classes 7, 8 and 9 are based upon a population which is considerably restricted, the increase in cor-

Table 2

Correlations of Officer Classification Test Scores with Criterion Measures at NTSch (Tactical Radar) Presented by Classes

Officer Classification Test Parts	The Correlation Coefficients Show the Relationship Between the Parts of the Officer Classification Test and Various Criteria of Success				
	Class 3 N = 178			Classes 7, 8, 9 N = 83	
	Theory Grade	Practical Grade	Final Grade	Achiev. Grade	Final Grade
1. Verbal	.16	.13	.18	.31	.32
2. Mechanical	-.06	.07	.04	.30	.19
3. Mathematical	.07	-.04	.01	.44	.49
4. Spatial	-.09	.01	-.02	.13	.08

relations is all the more significant. Again it appears that the increase can be attributed to the changed criteria of success in training.

CIC Aptitude Test. The effect of revising both the predictive measures and the criterion measures is demonstrated by the results presented in Table 3. Correlations presented in this table indicate that the CIC Aptitude Test is effective in predicting success in training of Officers at NTSch (Tactical Radar). Whereas, the correlations of the original Tactical Radar Aptitude Test Battery (Table 1) fluctuated considerably from class to class, these correlations represent a stable picture throughout, for the parts of the test as well as for the total score. It is also interesting to note that the test correlates as well with the final grade as

Table 3

Relationship Between CIC Aptitude Test Scores and Criteria of Success for Classes 7, 8 and 9 at NTSch (Tactical Radar)

CIC Aptitude Test Scores	The Correlation Coefficients Show by Classes the Relationship Between the Parts and Total Score of the CIC Aptitude Test and the Criterion Measures					
	Class 7 N = 66		Class 8 N = 123		Class 9 N = 117	
	Achiev. Test	Final Grade	Achiev. Test	Final Grade	Achiev. Test	Final Grade
Part 1	.49	.39	.37	.45	.42	.43
Part 2	.55	.46	.38	.51	.56	.56
Part 3	.48	.31	.40	.51	.47	.39
Total	.61	.45	.45	.55	.57	.56

with the achievement test grade. The correlations with final grade obtained for classes 8 and 9 are probably about as high as could be expected for an officer training prediction study.

Interrelationship of Criterion Measures

In a prediction study, one of the crucial factors is the nature of the criterion against which predictive indices are to be correlated. Some light on the nature of the criterion is shed by the results presented in Tables 4 and 5. The striking fact in Table 4 is the low correlation be-

Table 4
Intercorrelations of Criterion Measures

	Class 5 N = 134		
	Achiev. Test	Theory Grade	Practical Grade
Achievement Test			
Theory Grade	.59		
Practical Grade	.11	.13	
Final Grade	.80	.89	.31

tween the “practical” grade and the “theory” grade and between the “practical” grade and the achievement test score. The correlation between achievement test scores and “theory” grades of .59 indicates that while these two measures are definitely related, they do measure somewhat different facets of the student’s knowledge.

It should be remembered that the “practical” grade consists of the rating which was made of the officer’s performance in CIC. While one would not expect a perfect correlation between such a measure and the officer’s performance in class work, it does not seem reasonable that the relationship should be as low as shown in this table. If knowledge about a subject contributes to performance, one would expect the relationship to be higher, since the rating purports to be an evaluation of the officer’s performance in the CIC.

The intercorrelations in the case of class 9 are presented in Table 5. These data indicate that elimination of the rating of performance in CIC resulted in a substantial increase in the intercorrelation of criterion measures. The correlation of .66 between the achievement test and the second month grade and of .69 between the first month and second month grades correspond more nearly to what one would expect the correlations between the different criteria of success to be. In interpreting the cor-

Table 5
Intercorrelations of Criterion Measures

	Class 9 N = 117		
	Achiev. Test	First Month Grade	Second Month Grade
Achievement Test			
First Month Grade	.78		
Second Month Grade	.66	.69	
Final Grade	.91	.93	.84

relations between the different criterion measures and final grades, it should be remembered that they are spuriously high due to the fact that each criterion measure makes up one-third of the final grade.

Reliabilities of Predictive and Criterion Measures

The reliability coefficients of the predictive and criterion measures are shown in Table 6. They were computed by the Kuder-Richardson method except for the "theory" grade and the "practical" grade. The reliability of the "theory" grade was computed by correlating weekly grades for the odd numbered weeks of the course with those for the even numbered weeks of the eight-week course. The reliability of the "prac-

Tabel 6
Reliabilities of Predictive and Criterion Measures

Officer Qualification Test	.92
Officer Classification Test	
1. Verbal	.92
2. Mechanical	.83
3. Mathematical	.78
4. Spatial	.81
Tactical Radar Aptitude Test	.90
1. Polar-Grid Coordinate	.89
2. Ratio Estimation	.85
3. Coordinate Reading	.67
Relative Movement Test	.62
CIC Aptitude Test	.92
1. Polar-Grid Coordinate	.85
2. Scale Reading	.85
3. Relative Movement	.82
Theory Grade	.79
Practical Grade	.77
Achievement Test	.85

tical" grade was obtained by correlating the ratings assigned by two different raters. The class for which this particular reliability coefficient was computed was unique in that two ratings were available for every officer in every trait. For the majority of classes only fragmentary information was available concerning any one officer. For this reason it seems justifiable to conclude that this reliability coefficient represents the upper level of reliability for the "practical" grade.

Discussion

The most significant finding of this study is the influence of the nature of the criterion upon the relationship between predictive indices and measures of success. The low correlations obtained with the early classes included in this study can in part be attributed to the fact that the criterion of success was not predictable by means of the types of tests used. Whether the criterion was appropriate for the type of course offered at NTSch (Tactical Radar) is, of course, a different question. However, the low relationship between "theory" and "practical" grades and the observations of the staff at NTSch (Tactical Radar) that rating procedures were not operating properly, lend credence to the belief that an improved criterion resulted from the introduction of objective course examinations and the final achievement examination. Certainly it can be said that the criterion in use with classes 7, 8 and 9 was predictable while the one in use with classes 1, 2 and 3 was not.

A second major finding is the fact that the success in training of officers who are candidates for a specialized operational billet aboard a combatant vessel can be predicted with considerable efficiency. Since Naval officers represent a select population, it might have been assumed that any individual who is qualified to be a Naval Officer could qualify for a specialized operational billet such as Combat Information Center Officer. The results obtained in this study indicate that there are important individual differences among Naval Officers and that for a specialized operational billet some officers are definitely better qualified than others. This emphasizes the importance of careful screening of candidates for specialized operational billets, such as tactical radar, as well as for highly technical billets such as engineering and technical radar.

A third fact which is evident in this study is the need for continuous refinement and improvement of predictive indices. The original tests and selection requirements used in this study represent good estimates of what was required to aid interviewing officers in selecting suitable candidates for tactical radar training. Results soon showed, however, that the Tactical Radar Aptitude Test and Relative Movement Test could be revised and improved. Revision of these tests, resulting in the con-

struction of the new CIC Aptitude Test, brought about substantial improvement in prediction coefficients. Results such as these underscore the need for continuous research if selection requirements are to remain valid.

A fourth significant fact is the outstanding quality of the officers sent from indoctrination and reserve midshipmen schools to NTSch (Tactical Radar). The high average scores made in the Officer Classification Test by the members of classes 7, 8 and 9 indicate that tactical radar officers are drawn from the upper 15 per cent of the Naval Officer population. This fact contributes in part to some of the low correlations which were obtained in this study. Since very few failures occurred in the training, it was not possible to correlate predictive indices with a "success-failure" criterion. If an unselected population of officers had been sent to NTSch (Tactical Radar), the obtained correlations would have been higher, but it would also have resulted in a larger failure rate and, in addition, the fleet would have received tactical radar officers of markedly lower caliber.

Summary

The purpose of this study was to investigate the validity of the techniques used to select officers for tactical radar training and to determine the suitability of the criteria of success employed at Naval Training School (Tactical Radar). In the main, it was found that:

1. The Officer Qualification Test, Tactical Radar Aptitude Test, and the Relative Movement Test did not correlate highly with the scholastic success of students renrolled in the early classes.
2. The verbal and mathematical parts of the Officer Classification Test and the CIC Aptitude Test which was a revision of the Tactical Radar Aptitude Test showed substantial correlations with scholastic success of later classes.
3. The increased magnitude of the correlations obtained with the latter tests is partially attributable to the refinement of the criteria of success by the introduction of objective course examinations and the use of a comprehensive final achievement test.

Received December 31, 1945.

Prediction of Achievement in Typewriting and Stenography in a Liberal Arts College

Dorothy M. Barrett

Hunter College of the City of New York

Mindful of the day when they must seek employment in a world for which an A.B. degree may be inadequate vocational preparation, many Hunter College students have been adding typewriting and stenography to their studies in the liberal arts. Because a number of these young women discovered after a considerable investment in time that they lacked the necessary aptitude, it seemed imperative to try to find the means of predicting in advance a student's probable degree of success or failure in these subjects.

The results obtained in this study could be used either in the selection of students for a course when more students register than can be admitted, or to counsel the individual student who is debating the advisability of studying typewriting and shorthand.

Procedure

A total of 96 students who had registered in the course in typewriting and 75 students who had registered for stenography took the tests. Administered after the students had signed into the courses but before they had begun class work, the tests included Bennett's Stenographic Aptitude Test, the Kuder Preference Record, the MacQuarrie Test for Mechanical Ability, the Minnesota Vocational Test for Clerical Workers, Strong's Vocational Interest Blank for Women, Thurstone's Vocational Interest Schedule, and the Turse Shorthand Aptitude Test.

Final grades in each course were taken as the criteria of success. The final grades were based on speed and accuracy as demonstrated in class periods and in tests administered at the end of the course. End term grades were in no way influenced by the aptitude and interest test scores inasmuch as no instructor knew the results of these tests.

Any student who earned an A or B in the course under discussion will be referred to in this study as good; any student who earned D or F will be called poor or a failure, even though D is a passing grade for the course. Because students who earned C seemed to be neither true failures nor true successes, they have been classified separately.

Results in Typewriting

In Table 1 are listed the tests which differentiated between good and poor typists. Also included in the table are the chances for getting an A or B, C, D or F for each of several ranges of scores for each test.

Table 1

Showing Distribution of Grades in Typewriting for Several Ranges of Scores for Each of Several Aptitude Tests

Tests	A or B	Grades	D or F	No.
		C Chances in 100		
Minnesota Vocational Test for				
Clerical Workers				
Number Comparison				
150 and over	83	13	4	(23)
100-149	61	11	28	(67)
Below 100	33	17	50	(6)
Name Comparison				
150 and over	81	11	8	(37)
130-149	64	9	27	(33)
Below 130	42	16	42	(26)
MacQuarrie Test for Mechanical Ability				
Tracing Test				
50 and over	83	3	14	(29)
0-49	57	15	28	(67)
Dotting Test				
21 and over	70	11	19	(74)
0-20	46	14	40	(22)
Pursuit Test				
22 and over	74	14	12	(43)
14-21	59	11	30	(46)
0-13	43	0	57	(7)
Turse Shorthand Aptitude Test				
Total Score				
420 and over	80	10	10	(30)
Below 420	58	12	30	(66)
Total Undifferentiated Group	65	11	24	(96)

Both parts of the Minnesota Vocational Test for Clerical Workers distinguished students who earned A or B grades from those students who earned C, D or F grades at the end of the term, a fact which would be anticipated by the work reported by Andrew and Paterson¹ and by

¹ Andrew, D. M., and Paterson, D. G. Measured characteristics of clerical workers. Bull. of Empl. Stab. Res. Inst., Univ. of Minn., 1934, Vol. III, No. 1, pp. 1-60.

Eriksen.² In fact, the test of number comparison differentiated between good and poor students more accurately than any of the other tests included in this study. The chances for an A or B were 83 in 100 for those students who scored 150 and over on the number comparison test with only four chances in 100 for failing. The prevailing chances for the group as a whole for an A or B were 65 in 100 with 24 chances in 100 for failing.

The pursuit, tracing and dotting parts of the MacQuarrie test also differentiated between good and poor typists. The other parts of the MacQuarrie test differentiated poorly or not at all. Although Turse makes no claim that his test predicts efficiency outside of the field of shorthand, the composite total score did differentiate to a fair degree between good and poor typists.

Table 2

Showing the Distribution of Grades in Typewriting for Students Selected by Successively Imposed Critical Scores

Test Scores	A or B	Grades	D or F	No.
		C Chances in 100		
Number Comparison—150 or over	83	13	4	(23)
Name Comparison—150 or over	75	15	10	(20)
Tracing—50 or over	85	0	15	(13)
Undifferentiated Group Remaining	57	13	30	(23)
Below 22 on Dotting and below 22 on Pursuit	23	12	65	(17)

For the group of 96 students taking typewriting, no relationship to grades was found for the two scores on the Bennett Stenographic Aptitude Test, the Commercial factor on the Thurstone Interest Schedule, the ratings for General Office Worker or for Stenographer-Secretary on the Strong Vocational Interest Blank for Women, the ratings for Clerical Interest on the Kuder Preference Record, nor for the remaining parts of the MacQuarrie test.

An attempt was made to increase the effectiveness of the predictions of grades in typewriting by combining the results of several of the tests. Table 2 shows the distribution of grades when the scores for the number comparison, name comparison, tracing, dotting and pursuit tests were used successively or in combination to differentiate between good and poor students. This particular combination of test scores proved to yield the best results.

² Eriksen, E. G., et al. A demonstration of individualized training methods for modern office workers. Bull. of Empl. Stab. Res. Inst., Univ. of Minn., 1934, Vol. III, No. 2, pp. 1-60.

We began by eliminating from further consideration the 23 students with scores of 150 or over for number comparison. Reference to Table 2 will show that these students had 83 chances in 100 for an A or B and only four chances in 100 for a D or F.

For the 73 students who remained of the original 96, we set a minimum score of 150 for name comparison, thereby identifying twenty more students. These students had 75 chances in 100 for an A or B and only 10 chances in 100 for a D or F.

Next, a minimum score of 50 on the tracing test was set, thereby picking out another 13 students whose chances for an A or B were 85 in 100 but whose chances for failure were 15 in 100.

At this point, 56 students had been eliminated from the group, leaving 40 students. From this group it was possible to sort out 17 students who had a score below 22 on both the dotting and pursuit parts of the MacQuarrie test. Their chances for success were only 23 in 100, and for failure were 65 in 100.

The remaining undifferentiated group of 23 students did not yield to further analysis in the effort to sort out the good from the poor students on the basis of test scores. These students can be identified in Table 2 as the undifferentiated group having intermediate chances for success and failure.

The author concluded that the administration of the Minnesota Vocational Test for Clerical Workers, and the tracing, dotting and pursuit parts of the MacQuarrie Mechanical Ability Test represented a brief but effective combination of tests which would provide a fairly good estimate of aptitude for typewriting as taught at Hunter College.

Results in Stenography

In Table 3 are listed the tests which differentiated between good and poor stenography students. Also listed in the table are the chances for getting an A or B, C, or D or F for each of several ranges of scores for each test, as well as for the total group undifferentiated by any test scores.

On the basis of the distribution of grades for the 75 students of stenography in this study, without reference to any test scores, a student had 72 chances in 100 for an A or B, 21 in 100 for a C, and seven in 100 for a D or F. Any test which yielded scores which resulted in greater chances for success or greater chances for failure than those which prevailed for the group as a whole was judged a useful measure.

The transcription scores on the Turse Shorthand Aptitude Test, the pursuit and blocks scores of the MacQuarrie Test for Mechanical Ability, and the number comparison scores for the Minnesota Vocational Test for Clerical Workers differentiated most clearly between those students who

Table 3

Showing Distribution of Grades in Stenography for Several Ranges of Scores for
Each of Several Tests

Tests	A or B	Grades	D or F	No.
		C Chances in 100		
Minnesota Vocational Test for				
Clerical Workers				
Number Comparison				
150 and over	85	8	7	(13)
110-149	72	23	5	(53)
0-109	56	33	11	(9)
MacQuarrie Test for Mechanical Ability				
Tapping				
48 and over	81	10	9	(21)
0-47	69	26	5	(54)
Dotting				
26 and over	81	19	0	(16)
0-26	70	22	8	(59)
Copying				
45 and over	75	25	0	(12)
25-44	76	16	8	(42)
0-24	67	24	9	(21)
Blocks				
14 and over	86	4	10	(21)
0-13	67	28	5	(54)
Pursuit				
24 and over	89	8	3	(26)
0-23	63	29	8	(49)
Bennett's Stenographic Aptitude Test				
Transcription				
110 and over	77	19	4	(53)
0-109	59	14	27	(22)
Spelling				
38 and over	78	17	5	(62)
	46	38	16	(13)
Turse Shorthand Aptitude Test				
Phonetic Association				
48 and over	78	18	4	(44)
0-47	65	26	9	(31)
Transcription				
70 and over	85	15	0	(20)
0-69	67	24	9	(55)
Kuder Preference Record				
Clerical				
55 and over	76	24	0	(37)
0-54	69	18	13	(38)
Total Undifferentiated Group	72	21	7	(75)

did well in stenography and those students who did only average or poor work. Scores on the tapping, dotting and copying parts of the MacQuarrie test, scores for transcription and for spelling on Bennett's Stenographic Aptitude Test, and clerical scores on the Kuder Preference Record gave predictions which were an improvement over those predictions which could be made with no test results, but were less effective than the first mentioned sets of scores. The actual effectiveness of each test can be read directly from Table 3.

The scores for the remaining parts of the Turse Stenographic Aptitude Test and the MacQuarrie Test for Mechanical Ability as well as the scores for Stenographer-Secretary on the Strong Vocational Interest Blank for Women and the scores for Commercial interest on the Thurstone Vocational Interest Schedule failed to show any significant relationship to grades in stenography for the group of students studied.

The effectiveness of the predictions of grades in stenography was improved by combining the results of several tests. Table 4 shows the results when the Turse transcription scores, the pursuit scores for the MacQuarrie test, the number comparison scores for the Minnesota test, and the phonetic association scores of the Turse stenographic test were used in combination.

Table 4

Showing the Distribution of Grades in Stenography for Students Selected by Successively Imposed Critical Scores

Test Scores	Grades			No.
	A or B	C Chances in 100	D or F	
Turse Transcription—70 or over or	89	8	3	(36)
Pursuit—24 or over				
Number Comparison—150 or over	80	20	0	(5)
Turse Association—48 or over	67	29	4	(21)
Remaining Group	33	50	17	(13)
Total Unclassified Group	72	21	7	(75)

From the use of a score of 70 or over on the Turse transcription test or a score of 24 or over on the pursuit test, 36 students were singled out who had 89 chances in 100 for an A or B and might therefore have been accepted for training at once. Of the remaining group of students, 5 had a score of 150 or over on the number comparison test. These students had 80 chances in 100 for an A or B and might also have been accepted for the course.

Next it was possible to single out of the students still remaining, a group of 21 students with 67 chances in 100 for success and with 33 chances in 100 for being only average or poor. Finally, 13 individuals remained who had limited chances for success and 67 chances in 100 to be only average or poor. Individuals in both of these last two groups might well have been warned that they would have to exert themselves to be able to compete successfully with the individuals with higher scores.

In conclusion, then, as a basis for advising students about to study stenography, the data seemed to warrant the use of the transcription and phonetic association tests of the Turse Shorthand Aptitude Test, the pursuit scores of the MacQuarrie Test for Mechanical Ability, and the number comparison test of the Minnesota Vocational Test for Clerical Workers.

Summary

Grades earned at the end of one term of stenography or typewriting studied at Hunter College of the City of New York were related to scores on a series of aptitude and interest tests in the case of 96 students taking typewriting and 75 students studying shorthand. Administered after the students had signed into the courses but before they had begun class work, the tests included Bennett's Stenographic Aptitude Test, the Kuder Preference Record, the MacQuarrie Test for Mechanical Ability, the Minnesota Vocational Test for Clerical Workers, Strong's Vocational Interest Blank for Women, Thurstone's Vocational Interest Schedule, and the Turse Shorthand Aptitude Test.

The number and name comparison scores from the Minnesota Vocational Test for Clerical Workers, the tracing, dotting and pursuit scores of the MacQuarrie Test for Mechanical Ability and the total scores from the Turse Shorthand Aptitude Test differentiated between good and poor typists. However, for practical purposes, the author concluded that the two scores from the Minnesota Vocational Test for Clerical Workers, and the tracing, dotting and pursuit scores of the MacQuarrie test provided satisfactory predictions for advising students interested in studying typewriting.

Of the considerable number of tests which differentiated between good and average stenography students, the pursuit scores from the MacQuarrie test, the number comparison scores from the Minnesota test, and the transcription and phonetic association scores on the Turse Shorthand Aptitude Test proved a combination which provided the maximum possible predictions on the basis of the test data reported in this study.

Received January 5, 1946.

Readability of Mixed Type Forms *

Miles A. Tinker and Donald G. Paterson

University of Minnesota

Some newspapers, in order to produce a supposedly high degree of reader attention, introduce a medley of typographical arrangements on the same page or in different sections of the same feature article. Newspaper editors refer to this as "change of pace."

A moderate degree of "change of pace" typographical arrangement may be the practice in most newspapers. Nevertheless it can be carried to an extreme in some cases. For instance, a feature story occupying about one-third of a page in the Sunday edition of a metropolitan paper was printed with the following variations in typography: ordinary Roman lower case, italics in both lower case and all-capitals, all-capitals, bold face in lower case, capitals and italics, different line widths, different type sizes and amounts of leading, and boxed-in material. Most variations appeared several times, each time for a phrase, a sentence or a paragraph. A sample is shown in Figure 1. The justification for this kind of printing practice should rest upon experimental evidence and not upon the views of a particular editor no matter how experienced the latter may be. In addition to achieving certain reactions from the reader, the "change of pace" should also receive reader approval and the text should not sacrifice readability.

The present study was undertaken to measure the readability of and reader preferences for two medley typographical arrangements in comparison with straight-forward lower case Roman type. Readability was measured in terms of speed of reading and preferences were determined in terms of judged legibility and judged pleasingness.

The reading material consisted of Forms A and B of the Chapman-Cook Speed of Reading Test. Although performance on Form B is in general equivalent to performance on Form A, a control group was introduced to check the equivalence. There were 30 paragraphs of 30 words each in each test form. Reading time allowed was $1\frac{3}{4}$ minutes on each form.

The test forms were printed in the following typographical arrangements: (1) Form A and Form B were printed in 7 point Excelsior news-

* Grateful acknowledgment is given to the Graduate School, University of Minnesota, for research grant to finance this study.

paper type with a 12 pica line width and one point leading on newsprint paper stock. Form B was also printed in medley arrangement No. 1 which involved the following typographical variations: (1) Ten point Roman lower case, 12 pica line width, 2 point leading: (2) Same as (1) with italic rather than Roman; (3) Seven point Roman lower case, 12 pica line width, 1 point leading: (4) Same as (3) but in bold face; (5) Same

HERE'S THE WAY AA WORKS

Alcoholics' Problems Disappear---Presto!

By **THE AA GROUP**

THIS IS THE WAY Alcoholics Anonymous works:

You want to quit drinking? All right! You've tried often but you haven't been able to lick the stuff. Is that right? O.K.

ARE YOU WILLING TO ADMIT YOU ARE LICKED, that John Barleycorn is too much for you to tackle?

ARE YOU PREPARED TO POCKET YOUR PRIDE and frankly and freely confess your own weakness?

Very well. Now we're getting somewhere. Now then—

You know there is a God, don't you? No, no, we're not talking about some particular, more or less dogmatic conception. We simply mean a Higher Power, something up there which makes the universe tick.

It doesn't matter what faith you profess, or once professed. Maybe your an agnostic. But you must, a matter of simple reasoning, realize that something—some power—put here.

★ ★ ★

Oh, you do. Well, do you admit that this Power is considerably bigger than you are?—that it could succeed where you'd fail?

WELL, WHY DON'T YOU LET THAT POWER TAKE OVER THIS JOB?

That, in a nutshell, is the method of Alcoholics Anonymous, the method by which scores of men and women in Minneapolis have conquered John Barleycorn and are daily helping others to lick the stuffing out of him.

It isn't psychology. It isn't religion either, except that it is formed of the same stuff of which religion is made.

IT IS A SIMPLE, PRACTICAL KIND OF SPIRITUALITY. That is the best description of it.

It goes far beyond mere self-analysis. These cured alcoholics say that self-analysis, no matter how honest, is not enough.

True, this is the beginning of your release. But once having frankly examined your own condition, forming a resolution to quit won't be of any lasting help.

★ ★ ★

In short, will power won't do it. The AA's insist that no matter how strong an alcoholic's resolve, he can't stay dry that way. The craving for liquor is a disease and beyond the permanent control of the will.

Plagues of wife and friends won't keep him dry. Anger and tears are useless. An alcoholic may lose his job, everything he has, and still he won't quit because he can't.

He must throw himself unreservedly upon the Power above.

THERE ARE PROBABLY as many conceptions of God in the Minneapolis AA group as there are members. Some, of course, cling to a conception which agrees with the faith in which they were brought up.

Others have naturalistic conceptions. One AA told me he thought of God whenever he looked at an electric light. But whatever the conceptions, they are real and immediate. These folk make of God a familiar.

A peculiar thing happens, the AA's tell me, once an alcoholic puts his problem in the hands of the spiritual power.

RIGHT AWAY IT BEGINS TO DISAPPEAR. IS ONE OF THOSE INVOLVED KNOTS. No matter how hard you

THIS IS THE LAST of three articles a Alcoholics

Anonymous. The group publishes a 350-page book which describes the system in detail. Minneapolis address of the organization is P.O. Box 594. Groups also are organized in St. Paul and Hibbing. Addresses are Box 3452 in St. Paul and Box 461 in Hibbing.

work at unraveling it, you succeed only in making it tighter. Then you touch the right rope and presto!—it falls apart.

The accumulation of grudges, hates and irritations which have been driving an alcoholic to drink simply vanish.

The AA's say it is amazing how something which formerly seemed like an overpowering reason to get good and drunk doesn't seem that way at all any more.

It may be anything—difficulty with your job, irritating relatives, a nagging mother-in-law, jealousy of someone more successful. Whatever the may have been, it washes out.

The AA's call these things their "resentments." Alcoholics frequently are abnormally sensitive people and they have a lot of them.

They examine them one by one. Inspired by their new spiritual attitude they sweep the remnants of them right out of their minds. If, by nursing these resentments during their alcoholic days, they did anyone an injury, they proceed to correct it by calling on the person injured and making amends.

FROM THIS POINT ON, THEY ARE IN THE HANDS OF THEIR SPIRITUAL FAMILIAR.

Many of them use a very practical method by praying, every morning, for 24 hours of dryness.

They reject everything which might possibly develop into a new

FIG. 1. Medley arrangement or "change of pace" newspaper layout which prompted the present study.

as (4) but with 10½ pica line width; (6) Same as (3) but in all-capitals rather than Roman; (7) Same as (6) but with 10½ pica line width; (8) Same as (3) but in all-capitals bold face; and (9) Same as (3) but with 11 pica line width and boxed in.

In addition, Form B was printed in medley arrangement No. 2 with the following variations: (1) Ten point Roman lower case, 12 pica line width, 2 point leading; (2) Seven point Roman lower case, 12 pica line

width, 1 point leading; (3) Same as (2) but in bold face; (4) Same as (2) but in all-capitals; (5) Same as (4) but in bold face; (6) Ten point Roman bold face, 9 pica line width, 2 point leading; (7) Same as (6) but in a 10½ pica line width; (8) Same as (6) but in italics; (9) Ten point all-capital italics, 9 pica line width, 2 point leading; (10) Ten point Roman bold face, 10 pica line width, 2 point leading, boxed in; and (11) Same as (10) but in italics (not bold face).

In medley arrangement No. 1 all the printing was in Excelsior newsprint type face except the bold face which was in Memphis. Each variation involved a phrase, a sentence, or one or two paragraphs with repetitions. In medley arrangement No. 2 there was greater variation in line widths and more frequent changes from one arrangement to another within a paragraph, and boxed in paragraphs were used.

Three groups of 94 college students each served as subjects. Group testing in the classroom situation was employed. In each group, Form A was the standard. Each subject read the standard form first, followed by Form B typographically identical (control) or in one of the two medley arrangements. The order of presenting the test forms was systematically varied. See Tinker and Paterson¹ for details of methodology. Analysis of the data will reveal the influence upon readability of the medley arrangements in comparison with the standard arrangement.

Data for the speed of reading measurements are given in Table 1. In Test Group I, the control group, the results show that 0.31 paragraph must be added to the mean score on Form B in each test group to establish equivalence of the two forms. The data for Test Group II reveal that medley arrangement No. 1 was read 1.48 paragraphs slower than the standard arrangement. Similarly, as shown in Test Group III, medley arrangement No. 2 was read 2.00 paragraphs slower than the standard. The corresponding percentage differences are 8.35 and 11.39 respectively. The figures in column 10 show that these differences are statistically significant. As a matter of fact, our studies² have revealed few non-optimal typographical situations which retard speed of reading by more than 8 per cent, and very few that retard reading rate by as much as 11 per cent. The medley arrangements of type, therefore, produce a severe adverse influence upon readability of newsprint.

There appear to be at least three factors which operate to retard rate of reading in the medley arrangements of type: (1) Our previous in-

¹ Tinker, M. A., and Paterson, D. G., Studies of typographical factors influencing speed of reading: XII. Methodological considerations. *J. appl. Psychol.*, 1936, 20, 132-145; also Paterson and Tinker, *How to make type readable*. New York: Harper and Bros., 1940 (can be obtained from the authors).

² Paterson and Tinker, *op. cit.*, 1940.

Table 1
Effect of Medley Typographical Arrangements on Reading Speed

Note: Differences given are for the mean score on Form A minus the mean score on Form B printed in the standard or in medley arrangements (1) and (2).* All of Form A and Form B of control group were in 7 point Excelsior type face on newsprint paper stock. In each test group $N = 94$ college students.

Test Group	Comparison	Mean	P.E. Dist.	P.E. Mean	Diff. Between Means in		P.E. Diff.	r	$\frac{D}{P.E. Diff.}$
					Para- graphs**	Per cent			
(1) I	A 7 pt., 12 pica, 1 pt. leading	17.32	2.90	.30	0.00	0.00	.00	.86	0.00
	B 7 pt., 12 pica, 1 pt. leading	17.01	2.66	.27					
II	A 7 pt., 12 pica, 1 pt. leading	17.70	2.65	.27	-1.48	8.35	.16	.82	9.35
	B Medley arrangement No. 1	15.91	2.35	.24					
III	A 7 pt., 12 pica, 1 pt. leading	17.55	2.61	.27	-2.00	11.39	.16	.82	12.85
	B Medley arrangement No. 2	15.24	2.19	.23					

* See text for specifications of medley typographical arrangement.

** The differences in column 6 are "corrected" by the amount of the difference between the mean scores of Form A and Form B of Test Group I which serves as control group. The "correction" amounts to 0.31 paragraph for each test group comparison.

vestigations have shown that text in all-capitals seriously retards speed of reading and that material printed in italics retards reading slightly. Furthermore, readers have a strong aversion to reading text in either all-capitals or italics. (2) The shorter line width coupled with the larger size of type (10 point) interfere with effective perceptual habits in reading. That is, there are so few words per line that the important role of peripheral vision in speeding up perception is much less effective. When the line width is optimal, peripheral (and less distinct) vision of words along the line to the right of the fixation point gives premonitions of meanings and also guides the eye to successive fixations along the line. To eliminate effective use of peripheral vision, therefore, retards rate of reading. (3) The constantly occurring changes in typography (i.e., ordinary lower case to bold face to all-capitals to boxed-in material, etc.) are probably distracting to continuous and evenly sustained attention to meanings. This would also tend to retard speed of reading.

For the preference study, the whole test of 30 paragraphs was mounted on cardboard (the standard, medley arrangement No. 1, and medley arrangement No. 2) and 181 readers ranked the specimens according to opinions of legibility and according to preference as to pleasingness. The results of the judgments for apparent legibility are shown in Table 2. The average ranks reveal that the standard arrangement (7 point,

Table 2

Uniform and Medley Newsprint Arrangements Ranked According to
181 Reader Opinions of Legibility *

Type Variation	Average Rank	S.D.	Rank Order
7 pt., 12 pica, 1 pt. leading	1.83	.93	1
Medley Arrangement (1)	1.98	.56	2
Medley Arrangement (2)	2.19	.87	3

* See text for specifications of medley typographical arrangement.

12 pica, 1 point leading) was judged most legible, medley arrangement No. 1 next, and medley arrangement No. 2 poorest. This is the same order as for readability measured in terms of speed of reading. Note, however, that the differences between mean ranks are not large.

The results for judgments of pleasingness are listed in Table 3. Here, medley arrangement No. 1 was considered most pleasing, the standard in 7 point type was ranked next, and medley arrangement No. 2 was least pleasing. Again the differences between the standard and medley arrangement No. 1 were not large, but medley arrangement No. 2 was well separated from the other two kinds of printing. For medley arrangement

Table 3

Uniform and Medley Newsprint Arrangements Ranked According to
181 Reader Opinions of Pleasingness *

Type Variation	Average Rank	S.D.	Rank Order
7 pt., 12 pica, 1 pt. leading	1.86	.89	2
Medley Arrangement (1)	1.70	.52	1
Medley Arrangement (2)	2.44	.80	3

* See text for specifications of medley typographical arrangement.

No. 1, therefore, the ranks for pleasingness do not agree entirely with the ranks for judged legibility nor for the speed of reading results.

Although in earlier studies we have found a few differences between preferences and readability,³ the judgments for legibility and for pleasingness have tended to agree to a marked degree.⁴ The discrepancies in this study are difficult to interpret. Apparently readers consider that some variety in typographical arrangement is desirable from the viewpoint of pleasingness even though they consider such variation to be somewhat less legible than uniform typography. The readers who expressed the preferences have been exposed daily to a considerable amount of typographical variation in the local newspapers. Familiarity with such typography may have developed either a tolerance to or a liking for medley arrangements.

Is the practice of introducing a medley of typographical arrangements in newspaper printing based upon sound principles? On the one hand we find that the medley arrangements severely retard speed of reading and presumably ease of reading. Readers consider the medley arrangements less legible than uniform typography. On the other hand, readers who have been exposed to the practice judge the milder degree of medley arrangement to be slightly more pleasing than uniform typography, but dislike severe degrees of "change of pace" arrangement. Added to this, there are other reader reactions not measured in this study which may be either favorable or unfavorable. In making his decision concerning the use of medley arrangements, the editor should balance the factors of poor readability plus readers' unfavorable opinions in regard to readability *versus* whatever advantages are known to accrue from their use. It may be difficult to compensate for a loss of 8 to 12 per cent in readability and adverse reader opinions on legibility by other alleged advantages which may or may not be present.

³ Paterson and Tinker, *op. cit.*, 1940.

⁴ Tinker, M. A., and Paterson, D. G., Reader preferences and typography. *J. appl. Psychol.*, 1942, 26, 38-40.

Summary

1. The purpose of this investigation was to determine the readability of mixed type forms.

2. The speed of reading 7 point Excelsior newsprint in a 12 pica line width with one point leading was compared with the speed of reading two medley arrangements of newspaper type.

3. Medley arrangement No. 1 was read 8.35 and medley arrangement No. 2 was read 11.39 per cent more slowly than the 7 point Excelsior in uniform arrangement. This amount of retardation is serious and is seldom shown in non-optimal typography.

4. The slower rate of reading the medley arrangements is apparently due to several factors: (a) the slower rate for reading text in all-capitals, in italics and in non-optimal line widths, and (b) the possible distraction produced by frequently shifting from one typographical arrangement to another.

5. Judged legibility was in line with readability measurements. The 7 point newsprint in uniform arrangement was judged most legible, medley arrangement No. 1 was next, and medley arrangement No. 2 was rated least legible.

6. Medley arrangement No. 1 was rated most pleasing, the uniform 7 point text was next and medley arrangement No. 2 was a poor third. The difference in average rank between the first two, however, was not large. Apparently these readers tended to consider some variation in typography as more pleasing even though they judged such variation to be less legible than uniform typography.

7. In deciding to employ a medley arrangement in newspaper printing, the editor should consider whether certain alleged advantages more than compensate for the severe loss in readability and the adverse opinions of readers.

Received December 14, 1945.

Recombination of Ideas in Creative Thinking *

Livingston Welch

Institute for Research in Clinical and Child Psychology, Hunter College

Creative thinking or imagination is rated by many of the standard projective tests, such as the Rorschach test and the Thematic Apperception test. L. L. Thurstone and J. J. O'Connor and others have, in fact, devised special tests for this ability. So many factors, however, are involved in creative thinking that it seems desirable to study one, at least, which may be essential to all types. In this study it has been assumed that the ability to readily recombine or reorganize ideas according to some specific pattern is essential to all types of creative thinking, whether it be painting a landscape, inventing some new scientific instrument, or composing a new advertisement.

The recombination of ideas *per se* is common to mental activity which in the strict sense of the word we would not call creative thinking. Ideas in dreams are recombined to form images and ideas that we have never seen or thought of before. The motivation may be explained in terms of wishfulfilling, but there is no set plan or scheme involved in these recombinations. Such mental activity we might call phantasy. On the other hand, the creative artist or scientist recombines ideas in an attempt to achieve some goal or to solve some problem.

In an attempt to observe the part that the ability to recombine ideas according to plan plays in creative thinking, a test was constructed in which the subject was obliged to recombine familiar ideas according to four different patterns. The test was then given to a group of college juniors and seniors and to a group of professional artists. We were not pretending to test artistic ability alone. We did assume, however, that one of the many characteristics which are essential to artistic ability is the ability to recombine ideas quickly according to a plan and that where this ability was found to be sadly lacking, creative thinking might be significantly limited. We were willing to admit that one artist might be considered much greater than another and still the first artist might be slower in recombining ideas. In such an instance the superiority of the first artist could be explained in terms of the many other characteristics which he possessed; yet, despite these differences, we did expect the pro-

* The author is indebted to Dr. Louis Long, Miss R. M. Thomas and Miss Lee Clarke for their aid in constructing and administering the test.

professional artists as a group to be more successful on this test than the average intelligent layman represented by the college group.

Procedure

The test was divided into four parts. The first three sub-tests made use of written material and the fourth made use of blocks. The total testing time was 26 minutes.

Part I. Instructions

Recombine the words of each group on the next page to make as many meaningful, grammatical sentences as possible. For example, here is a group of ten words.

MEN SKY IS FIGHT THAT THE SLOW BRIGHT OF FOR
which can be recombined into the following sentences:

Men fight for the sky.
The sky is bright.
The fight is slow.
Etc.

You will receive as much credit for a *short* sentence as for a long one. Your sentences do not have to be artistic, but they must be grammatical. There must be at least a subject and a predicate. You will receive credit for a sentence which is only slightly different from another. A word from the group can be used only once in the same sentence, but it may be used any number of times in other sentences. Only use words from the group that you are examining at the time. You may skip from one group to another, if you like.

There are ten of these groups and you have only ten minutes in which to complete the test. Are there any questions? . . . Do not turn the page until the examiner says "Start."

The following are the ten groups of Part I.

1. DOG TREE CLIMBS RUNS THOSE A SMOOTH GOOD
BY WITH
2. CITY JOHN BUILT STOOD A THAT LARGE STRONG
FROM OF
3. CAR FENCE TRAVELS WAS THIS THAT BIG COOL
FOR BY
4. SEA WOMAN MOVE COULD THESE THE GREEN
ROUGH WITH OF
5. DEN LION ATE IS BIG DEEP THESE THE OF BY
6. HOUSE CHILD LEFT HAS BLUE FRIGHTENED THE A
FOR BY
7. LEMON WIFE COOKS FINDS THAT SOFT ROUND
WITH FROM
8. POTATOES MAID CUT ONCE SMALL HOT THESE A
OF FOR
9. FISH BOY WAITS CATCHES THE A LONG COLD BY
FROM
10. SLOWLY THE GOLDEN LIGHT THAT RESTED UPON
THEM MOVED AWAY.

Part II. Instructions

Make as many letters as possible using no more and no less than three straight lines. For example, the letter A is made with three straight lines, two

slanting downward and one across. You will be given no credit for the letter A, since it is an example.

Make as many letters as possible, using no more and no less than two straight lines.

Make as many letters as possible, using no more and no less than one straight line and one semi-circle.

The time limit is three minutes.

Part III. Instructions

On the next page you will be given a list of twenty words which you are to connect into a story. You must be certain to use the words in the order in which they appear on the list. If the first word on the list is "house" and the second word is "tree," you must first make use of the word "house" in your story and then make use of the word "tree." You must not skip any of the words.

Your story must be grammatical and logically related.¹ It must have a beginning and an end. You will be rated on the number of words you make use of in the time allotted. Write as fast as you can and underline each of the twenty words as you use it.

The time limit is three minutes.

The words used in this test were:

STAIRS OCEAN CHEMISTRY SONG TEST MOUNTAIN
BUBBLE DOG LEMON PICTURE POST BLANKET VIOLIN
LAMP NIGHTMARE STEAM LEG WINDOW SWAMP STAMP.
(The words were given in this order.)

Part IV. Instructions

The object of this test is to construct out of ten blocks on each trial, ■■ many pieces of furniture or home furnishings as possible. The piece of furniture you construct must fit properly. It must be symmetrical and be recognizable as ■ piece of furniture. Do not attempt to be futuristic. Use conventional forms. You must use a minimum of two blocks to construct a piece of furniture. You can use the same block over again to make another piece of furniture. You can make as many of the same type of furniture as you like. You will receive full credit for the same type that is only *slightly* different from another.

You have only ten minutes to complete this test. There are five trials. Hence, you have only two minutes for each trial.

In Figure 1, the forms of the ten blocks of one of the trays are presented. The blocks used in all five trials were similar geometric shapes selected

¹ The word "grammatical" as used in the instructions of Part I and Part III of this test simply means adhering to the standard rules of grammar. If the subject writes a sentence free of any errors of grammar, he obtains full credit for this sentence. The subject is given one small liberty,—the omission of the article before the subject or object, e.g. "Boy meets girl."

The phrase "logically related" in the instructions for Part III concerns the continuity of the story which must have ■ beginning and an end. An example of what is meant by logically related would be the following sentence, "The *stairs* led down to the *ocean*." An example of a lack of logical relationship would be, "I walk down the *stairs*. Last summer I took an *ocean* voyage. John studies *chemistry*." These three sentences are not descriptions of the *same* event. As long as the sentences are parts of the description of the *same* event, the subject obtains full credit.

from a box of playing blocks. On each trial the blocks were presented to the subject on a piece of cardboard with each shape outlined so that the positions of the blocks were standardized.

A record was kept of all of the combinations of blocks for which credit was given.

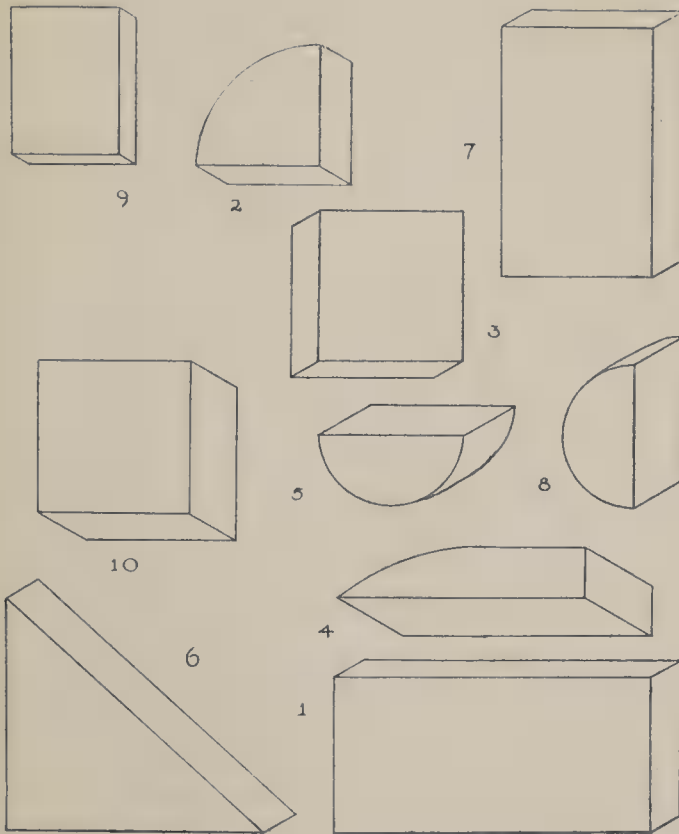


FIG. 1. The blocks used on one of the five trials of Part IV.

Subjects

The test was given to a total of 78 subjects, 48 college students and 30 professional artists. The students ranged in age from 18 to 29 with a mean age of 20. The artists ranged in age from 22 to 56 with a mean age of 37.

Results

The scores of the college group ranged from 23 to 56 with a mean of 37.6 and a standard deviation of 7.00, while the scores of the artist group ranged from 39 to 89 with a mean of 60.5 and a standard deviation of 12.26. The difference between the means was statistically reliable: ($D/\sigma = 9.3$).

The types of artists are classified as follows: 1 art director, 4 art teachers, 4 cartoonists, 2 fashion artists, 1 furniture designer, 5 illustrators 5 layout artists, 1 medical artist, 1 mural painter, 5 portrait and landscape painters, and 1 sculptor.

Table 1
The Means for All Four Parts of the Test

Parts	College Group <i>N</i> = 48		Artist Group <i>N</i> = 30		Critical Ratio between Means
	Mean	S.D.	Mean	S.D.	
I	18.0	4.24	17.7	7.15	0.2
II	6.7	1.83	12.5	1.93	13.2
III	9.1	3.15	11.4	4.09	1.1
IV	3.4	2.68	18.4	7.78	10.2

The Wonderlic Personnel Test was given to 48 of the college students and it was found that the correlation between the scores on this test and the test for the recombination of ideas was .27. No significant correlation was found between performance on the test for the recombination of ideas and chronological age. This indicates that the superior performance of the artists could not be explained in terms of their added years of experience.

There is no adequate way of rating artistic ability as in the case of academic achievement. We were fortunate enough, however, to have the opportunity of testing five commercial artists from the same advertising company and were able to compare the scores on the test with the company's opinion of the creativity of these subjects. The layout artist, who was considered the most imaginative, received the highest score (88), while the two fashion artists who were considered the most unimaginative received the two lowest scores (44 and 46).

The one furniture designer who took this test received a score of 47, which was low for the artist group. On Part IV of the test, which has to do with the construction of furniture, this person only made a score of 15. (The highest score was 35 points.)

Summary and Conclusions

In this study we have assumed that one essential element of any type of creative thinking is the ability to recombine ideas readily according to a pattern or plan. It is recognized that the creative artist has many highly developed abilities, but in this experiment only his ability to recombine ideas efficiently and quickly was tested. The test was divided

into four parts and the materials used were as familiar to the layman as they were to the artist. For example, the subject was required to recombine familiar words into different sentences and to recombine blocks into symmetrical pieces of furniture. No skill or training is needed for this test that would not be well developed in grammar school.

The test was given to 30 professional artists and to 48 college students. The mean score of the artist group was much higher than that of the college group.

Part II (making letters out of certain types of lines) and Part IV (making pieces of furniture out of blocks) differentiated the professional artists from the non-artists. Part I (composing sentences from certain word groups) and Part III (writing a story using a list of dissociated words) failed to differentiate significantly between the two groups.

The correlation between scores on this test and scores on the Wonderlic Personnel Test was very low. Moreover, chronological age did not appear to affect the scores.

Received January 14, 1946.

Questionnaire and Interview in Neuropsychiatric Screening *

Daniel H. Harris

Veterans Administration, New York City

The combination of a "neurotic" questionnaire plus individual appraisal in the weeding out of potential neuropsychiatric casualties from the flood of incoming recruits and inductees showed gratifying results during the last couple of years of World War II.

In the early days and months after Pearl Harbor, the NP screening of recruits by interview alone resulted in fantastic overworking of the few psychiatrists and psychologists available at the time to the armed services. Due to this overwork, it is likely that many misfits slipped through.

On the other hand, the procedure of rejecting men solely on the basis of a questionnaire does not appear to have been adopted at any time. However, the use of such an instrument as a preliminary coarse filter to separate out those to be interviewed, of whom a fraction will be finally rejected, has been found to be effective and time-saving (1), (3), (4).

In the writer's opinion, this fact has large possible implications for vocational and any other kind of selection involving large groups, in addition to its demonstrated value for military selection in peace or war.

Procedure

The present findings are based on the use in the manner indicated of an instrument developed at the Newport Naval Training Station. It consists of two parts: a 32-item condensation of the Cornell Selectee Index, Form N; and the Brown University Personal Inventory, Format C. Neither of these inventories has been released for general use, but results based on one or both have been published or referred to (1), (2), (3), (4). There have been no previously reported results based on Construction Battalion (Seabee) subjects.

Between Feb. 1 and May 2, 1945, a total of 2,081 Seabee recruits were received for training at the Naval Construction Training Center at Davisville, R. I. The training period was of ten weeks' duration.

■ The opinions and assertions contained in this paper are the private ones of the writer and are not to be construed as official or as reflecting the views of the Navy Department or of the Naval Service at large.

On arrival at the Training Center, each recruit filled out the two-part questionnaire. Administration was in groups of 50 to 100. After simple oral directions, it was usually completed in five or six minutes, and scoring by hand stencil took about 20 seconds per paper. Enlisted personnel aided in the administration and did the scoring, under supervision of the psychologist.

Using the "cutting" scores as used at Newport, 297 men were screened out by the questionnaire as possible NP casualties; the other 1,784 recruits were passed through by the questionnaire and went right on to duty without interview.

Each of the 297 men screened out was interviewed very briefly by the psychologist, usually within an hour or two of the questionnaire administration. These interviews averaged not over three minutes apiece; following which the man was: Sent on to full duty ($N = 203$); or Referred to the psychiatrist ($N = 52$); or Sent to trial duty ($N = 42$).

Of the 203 men sent to full duty, three were later referred to the NP department by their commanding officers during their ten-week training period. They were seen by the psychiatrist, whose ultimate disposition was: NP discharge from the Navy ($N = 2$); and Return to duty ($N = 1$).

Of the 52 men referred to the psychiatrist, four received medical or surgical discharges from the service before any NP disposition could be made. The psychiatrist made the following disposition of the remaining 48 men: NP discharge from the Navy ($N = 37$); and Return to duty ($N = 11$).

The 42 men sent to trial duty were called back for a brief re-interview after two or three weeks. This was generally even shorter than the first interview. Usually a written paragraph of appraisal of his adjustment to military service so far was available for each man, from his chief petty officer. In two cases the man had received a medical or surgical discharge before the re-interview could take place. Of the remaining 40 men, 26 were sent back to full duty following the re-interview, and none of these came again to the attention of the NP department. The other 14 were referred to the psychiatrist, whose ultimate disposition was: NP discharge from the Navy ($N = 7$); NP transfer to Naval Hospital ($N = 1$); and Return to duty ($N = 6$).

Of the 1,784 men who were not screened out by the questionnaire and so received no further NP attention on arrival, 23 were later referred to the NP department by their commanding officers during their training period. Their ultimate disposition by the psychiatrist was as follows: NP discharge from the Navy ($N = 15$); NP transfer to Naval Hospital ($N = 1$); and Return to duty ($N = 7$).

The initial and final disposition of all of the 2,081 recruits is shown in Table 1.

Table 1
Detailed Disposition of 2,081 Recruits

		Disposition by Psychiatrist				Medical or Surgical Discharge before NP Disposition
		Remained on Duty	Return to Duty	NP Dis- charge from Navy	NP Trans- fer to Naval Hospital	
Screened out by Questionnaire: 297	By Psychologist { Sent to Duty 203 Sent to Trial Duty 42 Sent to Psychiatrist 52	200	1*	2*		
		26	6	7	1	2
			11	37		4
		226	18	46	1	
Passed through by Questionnaire: 1,784			244		47	
		1761	7*	15*	1*	
	Total:		1768	61	2	63

* Indicates referrals to NP department by commanding officers during training.

Discussion

It is to be noted that the total number of "false positives"—i. e., men screened out by the questionnaire but immediately or eventually found fit for duty—was 244. This is 11.7% of the total number of recruits. At Newport (4) with general service personnel, the "false positive" rate was about 25%. It is probable that the difference is accounted for by the differing group characteristics of the two populations. Seabee inductees were in general older, occupationally more skilled, and undoubtedly differed in other, at present, non-demonstrable ways from the general service inductees taken in at a Naval Training Station.

Also worth mentioning are what might be called the "false negatives"—i.e., those who were passed through by the questionnaire but were later adjudged to be NP casualties after referral by commanding officers during training. As shown above, these numbered 16. This comes to 0.9% of the 1,784 men passed through by the questionnaire, and constitutes 25% of the total number of NP rejections.

As mentioned previously, the "cutting" scores here used were those used at Newport. They were: a score of 9 or more on the Selectee Index condensation, and/or a score of 1 or more on the Personal Inventory Format C. Analysis of the data to see what would have been the effect on the "false positive" and "false negative" rates of lowering the cutting score on the Selectee Index reveals that the following would have happened, with cutting scores ranging from the used score of 9 down to a score of 3:

Table 3

Cutting Score	False Positive Rate	False Negative Rate	
		% of Men Passed by Questionnaire	% of Total NP Rejections
9	11.7	0.9	25.4
8	14.0	0.8	23.8
7	15.1	0.7	20.6
6	18.6	0.7	20.6
5	23.5	0.6	17.5
4	32.9	0.5	14.3
3	47.8	0.4	12.7

It can be seen that the "false negative" rate does not go down as fast as the "false positive" rate goes up. However, in a situation where one is solely interested in cutting down the "false negatives" without caring how many perfectly good "false positives" he lost in the process, the

cutting score could be set so low that "false negatives" might be practically eliminated. This can be done only when there is a rather extravagant surplus of available manpower; but in this way it might be possible to eliminate practically all potential NP casualties by a completely automatic procedure involving no interviewing.

Summary

1. Of 2,081 Seabee recruits, 63 became NP casualties during their ten weeks' training period.

2. Forty-seven (75%) of these NP casualties were screened out by a 5-minute group questionnaire administered on arrival at the Training Center.

3. Of 203 men screened out by the questionnaire as possible NP casualties but sent to duty the same day by the psychologist after brief interview, 2 became NP casualties during training.

4. The combination of screening questionnaire and interview should be valuable for almost any kind of selection involving large groups.

5. By setting the "cutting" score low enough it may be possible under some conditions to select effectively by means of a questionnaire alone, without interview.

Received December 6, 1945.

References

1. Hunt, W. A. Clinical psychology in the navy. *J. clin. Psychol.*, 1945, 1, 99-104.
2. Weider, A., Mittelman, B., Wechsler, D., and Wolff, H. G. The Cornell selectee index: a method for quick testing of selectees for the armed forces. *J. Amer. Med. Assoc.*, 1944, 124, 224-228.
3. Weider, A., Mittelman, B., Wechsler, D., and Wolff, H. G. The Cornell selectee index: short form to be used at induction, at reception and during hospitalization. *New York Hosp., Cornell Univ. Med. Coll., and Bellevue Hosp.* (no date), 6 pp., mimeographed.
4. Wittson, C. L., and Hunt, W. A. Three years of naval selection. *War Med.*, 1945, 7, 218-221.

Restandardization of the Revised Beta Examination to Yield the Wechsler Type of IQ *

Robert M. Lindner

Haarlem Lodge, Catonsville, Maryland,

and

Milton Gurvitz

Hillside Hospital, Queens, New York

The Army Group Examination Beta developed during World War I has led to a considerable number of revisions and similar tests designed to permit measurement of illiterates, of persons who do not speak or read English, and of other individuals for whom a verbal test is not considered suitable. Perhaps the most important revision of this test is that by Kellogg and Morton¹ published in 1934 and entitled "Revised Beta Examination."

For many years this test has been used extensively. Its most common application apparently has been in penal institutions where it has been found useful for purposes of initial classification of committed persons. The scores have been found significant in relation to the psychiatric, educational, and vocational adjustment of persons in these institutions. A second major use of the test has been in selection and classification of employees in mass industries. The publisher reports that the test is usually sold in large quantities to institutions and large industries, although it is apparently used quite generally in small quantities for a variety of educational, vocational, and counseling purposes.

Because of its general usefulness, the authors were interested in improving the administration and standardization of the test. This opportunity arose because of the large number of cases which was available to them at the United States Federal Penitentiary at Lewisburg, Pennsylvania. Information was available concerning the subjects used which permitted a standardization to be made according to modern methods of equating the sample to the general population. A summary of the sampling procedure will be discussed below.

* This paper is based upon a more extensive unpublished report of the research. Dr. Harold Seashore of The Psychological Corporation assisted materially in its preparation. This is a "prior publication," authors paying costs.

¹ Kellogg, C. E., and Morton, N. W. Revised beta examination. The Psychological Corporation, 1935. Also see revised beta examination. *Personnel J.*, 1934, 13, 98-99.

In addition to planning a general restandardization there arose the question of converting the scoring and interpretation of the test into a more useful form. It was decided to follow the general method developed by Wechsler² in the standardization of the Wechsler-Bellevue Intelligence Scale.

The main features of the Wechsler type of scoring and standardization are, first, that each of the subtests is converted into scaled scores so that a profile is secured of the subtests and, second, that the computation of the IQ takes cognizance of the fact that mental ability as measured by the test declines with age after a peak of development in the early twenties.

Changes in Administration and Scoring of the Subtests

Minor changes have been made in the administration of some of the subtests, and certain of the scoring procedures for the subtests have been made more explicit and objective. There has been no change in the content of the test. These changes are of no interest at this moment; they will be presented in the revised manual which the publisher is making available.

Weighted Scores for the Subtests

At the present time the various subtests of the Beta Examination contribute differentially to the total and apparently there has been no demonstration that this is the optimum weighting. It can be assumed that each test is as good as any of the others. While strictly speaking this is not true, it is probably true enough to make any further attempt at refinement unwarranted. The authors therefore decided that the scoring should be arranged so that each subtest would contribute equally to the total score. The plan has two advantages. It allows an examiner to prorate a score when for some reason or another one or two of the subtests have to be omitted. In addition it may provide a valuable clinical tool to be used in screening out psychiatric deviates who frequently express their personalities by unequal performance on a battery of six subtests. Wechsler has presented considerable evidence along this line for his test and it is hoped that when this new standardization becomes more widely used, similar information regarding "scatter" can be developed for the Revised Beta Examination.

Each subtest is made to fit a 20-point scale with a mean of 10 and a standard deviation of 3. Using Hull's method,³ each raw point score of

² Wechsler, D. *The measurement of adult intelligence*. Baltimore: The Williams & Wilkins Co., 1941.

³ Hull, C. L. *Aptitude testing*. Yonkers, N. Y.: World Book Co., 1928, p. 397 ff.

each subtest is equated to the new scale. Tables of weighted scores are provided in the new manual for the test. This weighting was accomplished by taking 1,006 heterogeneous test papers that were available. Several other criterion test scores were available, and the educational level of these cases was known. When these criteria were correlated with the raw scores and then with the newly designed weighted scores, there were only small changes in the size of the coefficient of correlation. Because these correlations may be of some general interest, they are presented in Table 1.

Table 1

Correlation Coefficients of Beta Raw and Weighted Scores
with Other Variables ($N = 1,006$)

	Beta Raw Score (Kellogg & Morton) <i>r</i>	Beta Weighted Score (Lindner & Gurvitz) <i>r</i>
U. S. Public Health Service Classification Test Weighted Score	.90	.85
U. S. Public Health Service Classification Test IQ	.86	.87
Last School Grade Completed	.61	.58
Stanford Achievement Test, Paragraph Meaning	.75	.67
Stanford Achievement Test, Word Meaning	.74	.70
Stanford Achievement Test, Arithmetic Reasoning	.72	.68
Chronological Age	-.30	-.34

The Selection of the Standardization Sample

The research is based upon an original collection of Beta scores on more than 2,000 cases. This sample could not be considered a proper one upon which to standardize a test without further inquiry into its composition. It was soon discovered, as Wechsler had found, that the mean weighted score declined steadily when the persons in the sample were grouped in five-year intervals of age. This demonstrated clearly that the Beta Examination was functioning much as Wechsler's test and that therefore a general application of his method of computing the IQ might prove desirable. Before proceeding, however, further refinements of the sample were necessary.

All psychotics and extremely physically handicapped individuals were removed. The standardization, furthermore, was limited to male adults. Negroes were removed from the population. It was found, for instance,

that age for age the negroes had an average of about one and one-half fewer years of education than whites. With these preliminary refinements, a sample was secured of 1,800 white male adult prisoners ranging from 17 to 70, including almost every previous occupation, and having origins in all the states east of the Mississippi River and most of the Western States.

The next decision was to select from this group individuals who would be distributed in an educational grouping similar to that shown by the 1940 Census. Furthermore, this distribution was to take cognizance of age. A sampling was done in such a way that within each age grouping of 5 or 10 years in range, the individuals would be distributed educationally in proportion to the distribution of white, male adults in these same age ranges in the 1940 Census. The third variable in the selection process was the socio-economic status of the individuals, which also was equated to agree with the report of socio-economic groups, by age, in the 1940 Census.

The details of this sampling process are too elaborate to justify presentation in this report. The protocols are available from the junior author and can be supplied to anyone interested in the method. The

Table 2

Means and Standard Deviations of Weighted Scores of Standardizing Sample by Age Groups, and Corrected for Education and Socio-Economic Status ($N = 1,225$)

Age	<i>N</i>	<i>M</i>	<i>SD</i>
16-19	85	65.0	12.1
20-24	220	66.9	11.9
25-29	195	65.0	12.9
30-34	197	62.1	14.1
35-39	200	58.7	14.9
40-44	90	55.3	15.9
45-49	83	52.0	16.8
50-54	80	48.7	18.2
55-59	75	45.6	19.4

combined technique of the whole process of taking into consideration age, education, and socio-economic status resulted in the sample described in Table 2. This table presents the number of cases and the means and standard deviations of the weighted scores of the final standardizing sample, by age groups, and as corrected for education and socio-economic status. It will be noted that the mean weighted score declines and that the standard deviation increases steadily with increasing age. The smoothness of these lines indicates the quality of the sampling procedure.

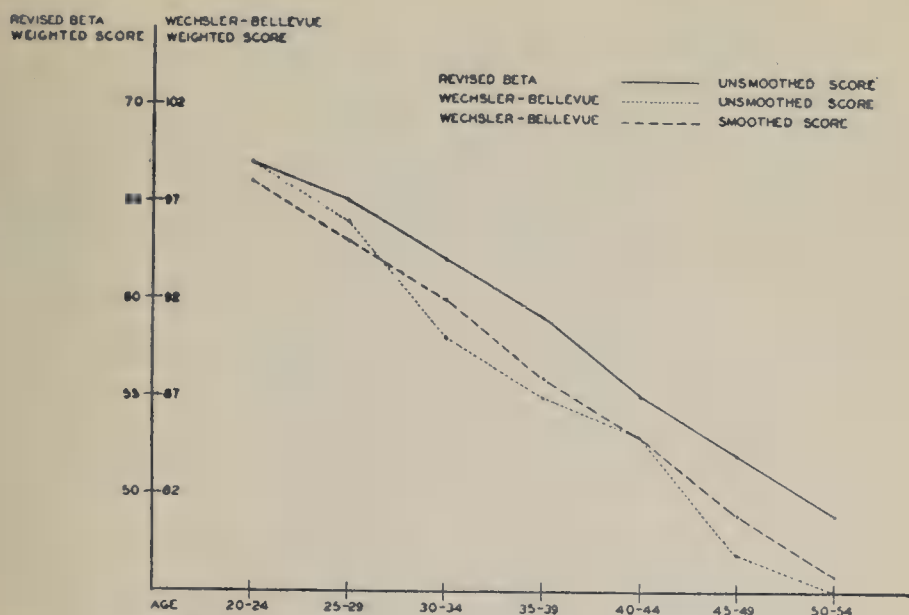


FIG. 1. Variation in means of weighted scores by 5-year age ranges.

A comparison of the Wechsler-Bellevue and the Revised Beta curves of mean score against age is shown in Figure 1.

The Derivation of Norms

Wechsler's method for calculating IQs has been followed in detail. The data for converting weighted scores for each subtest to IQs for each age range are provided in the new manual and need not be repeated here. Table 3 presents evidence that the conversion of weighted scores to IQs by the new table has accomplished the purpose desired. Observe that both the mean IQ and the standard deviation of the IQ are relatively constant from age range to age range.

Table 3

Means and Standard Deviations of the IQs for Each Age Range of the Beta Standardizing Group

Age	Mean	SD
20-24	100.1	14.8
25-29	99.8	13.7
30-34	99.8	14.4
35-39	99.7	14.9
40-44	100.8	14.7
45-49	99.8	14.2
50-54	99.4	14.8

Correlation of Beta and Wechsler-Bellevue IQs

One hundred and sixty-eight cases were tested with the Revised Beta Examination and with the Wechsler-Bellevue. The new scoring procedure resulted in a coefficient of correlation of .92 between IQs on the two tests. The mean difference between the IQ scores, irrespective of sign, is 7.2. There is a general tendency for the Revised Beta Examination to result in somewhat lower IQs for the same individuals, a tendency which can be regarded as desirable since it is known that the Wechsler-Bellevue is not so discriminating at the lower levels of ability. It is at the lower levels that Beta is particularly useful. On the other hand, the Beta Examination is not suitable for measuring IQs above about 120 or 125. It is also likely that both tests lose some discriminative sensitivity above the age of 40.

Figure 2 is presented to compare Beta and Wechsler IQs at different IQ levels at different ages. Each curve plots the IQ assigned to persons of a given age who have a weighted score which would yield the stated IQ at age 20-24. For instance, a person with a Beta IQ of 65 at age 20-24 would have a Beta IQ of 92 if he were age 50-54. The same person would have a Wechsler IQ at age 50-54 of 81.

One should remember that the IQ tables for these tests are constructed so as to yield an average IQ of 100 for each age group. The steepness of the curve in Figure 2 then indicates the correction for age which was necessary to make the IQs equal at different ages. Several observations can be made from these figures.

1. In general, the Beta test tends to require a greater correction for age as shown by greater steepness in most of the curves.
2. Both tests need more correction for age at lower IQ levels as indicated by the greater steepness of the curves for lower IQs.
3. The Beta IQ and Wechsler IQ curves are more discrepant at the lower IQs. No one knows whether the Beta or Wechsler data are better descriptions of the effect of senescence on ability. The general trends for the two tests are the same and the differences are doubtless due to the content of the tests and perhaps to the method of administration.
4. If an older person is assigned a Beta IQ of 100 and a Wechsler IQ of 100, the difference in actual test performances from that of 20-year olds is more on the Beta than on the Wechsler.

Sex and Race Differences

The question of whether the newly devised norms are suitable for females can only be answered on a theoretical basis as no women subjects were available to the authors for the standardization. The Beta has been

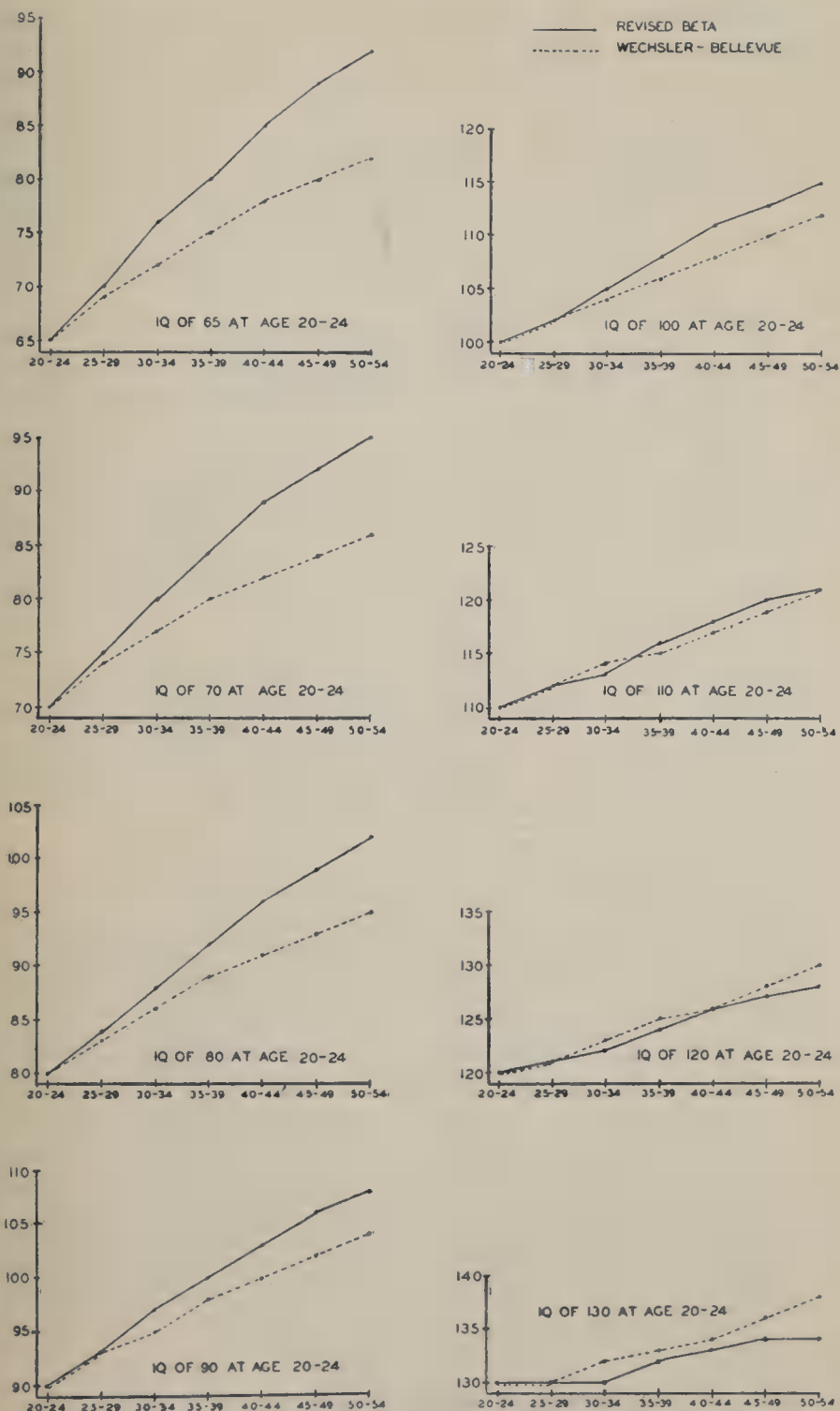


FIG. 2. IQ variation with age for a weighted score giving a specific IQ at age 20-24.

used for testing women as well as men for many years and the authors know of no report indicating that different norms were required for different sexes. A study of the census data indicates that there is no appreciable difference between the educational status achieved by native white males and native white females. Furthermore, there is no other intelligence test that we know of that uses a separate scale for men and women. If the test is to be used in a large-scale program where separate norms can be obtained, it might be desirable for someone to make an experimental study of this problem.

The authors have considerable evidence at hand to indicate that the difference in performance on the Beta Examination between whites and nonwhites is about roughly proportional to the difference in mean number of grades completed. This, of course, is not adequate evidence to suggest that the norms of the Revised Beta would be applicable to nonwhites. They made several other attempts to find a way of extending the standardization to make meaningful scores for negroes. This plan was abandoned for several reasons, but primarily because the sampling of negroes produced norms that seemed to exaggerate the differences between urban northern-born negroes, and rural southern negroes.

Psychologists should be thoroughly acquainted with the literature of the effects of culture and education on tests, and it is assumed they will have this background information when they are evaluating IQs on this test or any test when individuals from atypical cultures are being measured.

Suitability to the General Population

This test was standardized on adult, white, male prisoners. Can such a standardizing population produce norms applicable to the general population? Several considerations appear. The standardization does not reflect a prison population when it utilizes a sampling based upon educational and socio-economic standards determined by the 1940 Census. New residents of the Federal Penitentiary at Lewisburg were tested within one week of their entrance into the institution, a fact of importance since experience shows that incarceration for a longer period of time can develop stereotyped modes of thought and expression. It should also be pointed out that this penitentiary does not receive many established criminals; it is an institution for adults who are considered improvable offenders. They are essentially not criminals but lawbreakers.

The Classification of Intelligence

IQs as calculated by this revision of the Revised Beta Examination must be recognized as relative indices of the degree of intelligence. IQs

determined by this method should always be labelled "Beta IQ" and not simply "IQ." In interpreting the score one should also be aware of the fact that a Beta IQ of 70 in an older person is different from a Beta IQ of 70 in a younger person with respect to the performance on which it is determined. The authors would like to propose a method of reporting whereby both the IQ and the weighted score will be interpreted. Tables 4 and 5 present these two modes of classification.

Table 4
IQ Classification *
(Based on Weighted Scores and Age)

IQ	Classification
129 and up	Very Superior
120-128	Superior
110-119	Above Average
90-109	Average
80- 89	Below Average
71- 79	Inferior
70 and below	Defective

* Classification system same as that used by Wechsler.

Table 5
Weighted Score Classification
(Without Regard to Age)

Weighted Score	Classification
90 and above	A
83-89	B
75-82	C+
59-74	C
61-58	C-
43-50	D
42 and below	E

Intercorrelations of Subtests

The intercorrelations of the subtests of the Revised Beta Examination as administered and scored by the procedure reported by the authors in the new test manual are given in Table 6.

Summary

1. The Revised Beta Examination has been restandardized to accomplish three purposes: (a) The administration and scoring procedures have been improved. (b) The sample of adults upon which new norms

Table 6

Intertest Correlations of the Subtests and Weighted Score ($N = 1,006$)

	Weighted Score	Maze	Digit Symbol	Error Recog- nition	Form- Board	Picture Com- pletion	Iden- tities
Weighted Score	—	.68	.86	.82	.75	.83	.78
Maze	.68	—	.62	.51	.52	.55	.54
Digit Symbol	.86	.62	—	.60	.57	.67	.72
Error Recognition	.82	.51	.60	—	.74	.76	.58
Form-Board	.75	.52	.57	.74	—	.62	.51
Picture Completion	.83	.55	.67	.76	.62	—	.56
Identities	.78	.54	.72	.58	.51	.56	—
Average	.79	.57	.67	.67	.62	.67	.62

are based has been selected to represent the 1940 Census with respect to education and socio-economic status within several age groups from twenty years and above. (c) The standardization permits the securing of Beta IQs which are similar in meaning (though not necessarily in size) to the IQs secured on the Wechsler-Bellevue Intelligence Scale.

2. The procedures for the selection of the sample are briefly reviewed, and the authors feel that they have standardized the test on a sample of white, male adults above the age of 20, which is representative of the general population.

3. The method of weighting the subtests and determining the IQs as worked out in this research resulted in average IQs and standard deviations of the IQs which are equivalent for various age ranges. The average weighted score showed a steady decline with age similar to that shown in Wechsler's research.

4. It is believed that the Revised Beta Examination, when administered and interpreted according to the procedures outlined in the new manual for this test, should prove to be an excellent group test for measuring the mental ability of adults. One of its most useful applications will be in the initial classification of persons who are committed to mental and penal institutions. It is also recommended for use with illiterates, with persons who do not speak or read English, or with others for whom a verbal test is not considered suitable. It can be considered a satisfactory verifying examination in connection with other more verbal paper group tests or individually administered mental examinations.

5. When reporting results on Revised Beta Examinations, administered and scored according to this new standardization, the proper term to use is "Beta IQ."

Received October 14, 1946.

Book Reviews

N O R C Interview Department. *Interviewing for N O R C*. Denver: National Opinion Research Center, 1945. Pp. + 154. \$2.00.

From the time of the first public opinion survey by a nation-wide interviewing staff in 1935, critics have pointed out that the results of such polls are frequently impaired by the bias, dishonesty or carelessness of interviewers. The National Opinion Research Center has recognized the importance of the training problem implied in this criticism, and this book, written by the N O R C Interview Department as a manual for the 200 part-time interviewers who work for N O R C in all parts of the country, represents a painstaking effort to solve this problem.

Throughout the volume an attempt is made to give the part-time interviewer a complete picture of the various stages a N O R C survey must go through before and after the actual interviewing job is done. The interviewer is not merely given detailed rules and suggestions, but he is told specifically how his errors may cause difficulty for the central staff at Denver or New York. The interviewer is also told quite frankly what criteria are used in rating his ability. For instance, performance on free-answer questions is described as one of the principal measures of ability. An interviewer who records a great variety of comments in language which seems to mirror his cross section appropriately is considered superior to the interviewer who reports only brief stereotyped remarks using his own word-patterns over and over again. By taking the part-time interviewer into their confidence in this manner, the authors have attempted to make the interviewer feel that he is an integral part of the organization.

Although the volume deals specifically with N O R C problems, some sections are of general interest to anyone who is engaged in opinion poll interviewing. This is especially true of the section called "How to Get a Good Interview." The authors have attempted the very difficult task of teaching an interviewing technique which is objective, standardized, free from interviewer bias, and at the same time, informal and capable of securing more than superficial responses. Interviewers are instructed to accept "don't knows" and "qualified answers" only after probing for a more definite answer. Standardized probing is to be done primarily by repeating the question verbatim. Important words can be given a stronger emphasis or phrases which do not change the question meaning, such as "On the whole" or "Well, in general" can be prefaced to the question.

Other sections of the book dealing with such problems as quota controls, factual data, rural interviewing and telegraphic polls do not offer much information of general interest, although the material is presented in a readable manner.

To the best knowledge of the reviewer, no other similar manual for opinion poll interviewers is now available to the public. N O R C is to be commended for filling this need and for doing a workmanlike job.

Philip H. Kriedt

University of Minnesota

Smith, B. L., Lasswell, H. D., and Casey, R. D. *Propaganda, communication, and public opinion*. Princeton: Princeton University Press, 1946. Pp. 435. \$5.00.

This book is an invaluable reference book for the research applied psychologist. In addition to four essays on the science of mass communication, there is an annotated bibliography of 2,558 titles classified under seven main headings with numerous sub-headings. A total of 150 major titles are identified as being of especial value for the scientific student.

This Reference Guide is a continuation of the work by the same authors which was published in 1935 under the title *Propaganda and Promotional Activities: An Annotated Bibliography*. The 1935 book listed about 4,500 titles. The present book adds some 2,500 titles which have appeared for the most part since 1935. This indicates the rapid growth of scientific interest in the analysis of propaganda and other forms of mass communication.

The applied psychologist will be especially interested in those titles classified under theory and measurement. A total of 321 titles are included in the former category and 263 in the latter.

Donald G. Paterson

University of Minnesota

Munroe, Ruth L. *Prediction of the adjustment and academic performance of college students by a modification of the Rorschach Method*. Applied Psychology Monographs, No. 7, September 1945. Pp. 104. \$1.25.

One of the facts about the Rorschach literature which is disturbing to those American psychologists who are interested but skeptical is the pitifully small number of studies which can really be called "validation" studies in any respectable sense, buried among a great mass of investigations whose titles would suggest that they were systematic studies of validity, but which turn out not to be so at all upon reading. It is,

therefore, gratifying to come across this work of Dr. Munroe's, which is a contribution both to college guidance and to the case for the Rorschach. She presents data from three successive freshman classes, totalling 348 girls, at Sarah Lawrence College. Her aim was to determine the efficiency of the Rorschach in predicting "adjustment" and academic achievement. The Rorschachs for the first 100 cases were given in the usual manner, but scored by Munroe's "Inspection Method," and 60 of them were administered by someone else and scored later by the author. During the second two years of study the Harrower Group Rorschach was given and scored also by the Inspection Method. The "validating" criteria include several studies utilizing short "blind" personality sketches and the method of correct matchings, previously reported in part; academic achievement as indicated by scholarship ratings of the Student Work Committee (letter grades are not given at Sarah Lawrence); adjustment ratings as indicated by faculty consultations and referrals to the psychiatrist; and a rating of maladjustment by the Student Work Committee.

Because of the nature of the criteria available, data are not presented in correlational form, but are expressed in terms of contingency tables. In one respect this turns out to be a decided advantage, since it brings out certain relationships which a Pearson r as ordinarily employed would possibly obscure. For example, the Rorschach used alone is about equally efficient in predicting academic success as the ACE used alone. Contingency coefficients are .43 and .36 respectively. Since there is practically no relation between the two, combining them improves prediction slightly, as shown by a contingency coefficient of .50. It is noted that the ACE is more effective than the Rorschach in predicting definitely *superior* work, whereas students doing poor work in spite of high ACE scores tend to have "maladjusted" Rorschachs. These are, of course, the results that one might expect theoretically.

The study is unfortunately marred by a few minor defects and deviations from perfect control which leave a way out for anyone who is adamantly skeptical about the Rorschach. For example, it would have been better had all of the tests been scored completely "blind," without any opportunity for clinical impressions to be formed in face-to-face contact. The author states that the results on those cases she tested personally were not "significantly" better than the others. The third freshman class might well have been excluded to increase the purity of design, since their records were filed and available to teachers before the year's close. Here, however, Dr. Munroe states that the results for the third year were not significantly better than those of the first two. One would like to have the data separately analysed for the middle year,

when the most rigid control was exercised. On the whole, however, the study could be a pretty good model for other Rorschachers, and it certainly is an important contribution to Rorschach literature.

Paul E. Meehl

University of Minnesota

Rapaport, David (with the collaboration of Gill, Merton, and Schafer, Roy). *Diagnostic psychological testing*. Chicago: The Year Book Publishers, Inc., 1946. Vols. I and II, pp. xxii + 1098. \$13.00.

These volumes summarize an extensive and systematic investigation of the differential diagnostic potentialities of a psychometric battery consisting of the Wechsler-Bellevue Adult Intelligence Test and the Babcock Deterioration Test as tests of intelligence; the Object Sorting Test and the Hanfmann-Kasanin Test as tests of concept formation; and the Word Association Test, the Rorschach Test, and the Thematic Apperception Test as tests of ideational content and personality. Fifty-four "normal" control cases, selected at random from the Kansas Highway Patrol, and 217 psychiatric cases, apparently from the Menninger Clinic, serve as subjects. On the basis of psychiatric and social history data, the control group is divided into three sections according to excellence of adjustment. Similar but more extensive data provide for the categorization of the clinical cases as "schizophrenic," "preschizophrenic," "paranoid condition," "depressive," or "neurotic," a finer classification being employed as the needs of analysis dictate.

A substantial portion of the text is devoted to development of a psychological rationale for each test, along with supplementary validation data for some of the tests. Considerable space is given to discussion of the general application of those of the tests which hitherto have been credited with only a specialized and restricted utility. Test by test statistical and clinical comparisons between the control and clinical subgroups reveal the diagnostic potentialities of the battery. The numerous psychometric patterns and indicators of diagnostic import are discussed and described in some detail. The appendices include a review of pertinent literature on each test as well as the test scores and history data for each subject.

This research encompasses much crucial but essentially unexplored or controversy ridden territory in clinical psychology. The authors are both free and ingenious in proposing and provisionally substantiating hypotheses in their development and discussion of test rationales. These hypotheses and rationales are so numerous and extensive that they cannot be discussed adequately in this brief review, however.

Prolivity in the text, lack of illustrative clarity in the graphs, and

occasional reliance on verbal argument when statistical argument might have been more lucid and forceful weakens the presentation at times. In some instances verbal recapitulation of statistical tables unduly lengthens the text. Limitation of the statistical argument largely to the "t" and "Chi Squared" tests of significance may have reduced somewhat the fecundity and practical utility of the analysis, possibly forcing the authors to resort, for purposes of argument, to the less rigorous clinical impressions and verbal analyses more often than was strictly necessary.

The authors present no summarizing statistics on the differential diagnostic accuracy of the battery as a whole. One of their aims was "to show how the . . . tests . . . were welded in (their) clinical work into a single diagnostic tool." This omission is thus glaring and disappointing, the more so because of the crucial character of such a summary.

These volumes possess redeeming features which outweigh their defects, however. The wealth of clinical information they contain as well as the stimulating speculations and informative discussions concerning test validation, problems encountered in the clinical use of the tests, and possible psychopathology underlying various forms of impairment in test performance should make them worthwhile aids to clinical psychologists. Both the clinical and the heuristic value of these books more than justifies their inclusion in class reading lists and clinic libraries, though they can hardly be considered as text material for other than advanced courses.

HOWARD F. HUNT

Stanford University

Hayes, Samuel P. *Vocational aptitude tests for the blind*. Perkins Institution and Massachusetts School for the Blind, Watertown 72, Massachusetts, 1946. Pp. 32. 25 cents.

In a small readable book written in non-technical language, Hayes briefly surveys the attempts to develop vocational aptitude tests for the blind. The approach is historical, and gives quiet testimony to the prominent part played by the author in the development of psychological tests for this group.

The scope of the treatment is indicated by the opening definition of vocational aptitude tests as those "designed to measure the special abilities needed for success in some specific occupation." Hayes intentionally excludes not only general intelligence and achievement tests which he has discussed in numerous other articles, but also personality "tests," which need little change for use with the blind except their transformation into braille. The survey encompasses psychological measurements under

the classifications of mazes and formboards, musical, manual and mechanical, and scholastic aptitude tests.

The book is not a test administrator's manual but a description of the research efforts made to date within the scope of the given definition. At appropriate points in the book the author makes thoughtful appraisals based upon his experiments and experiences.

The dearth of research in this field is indicated by the inclusion of twenty titles in the bibliography. Fifteen of these are specifically on the blind. Only two of the articles report efforts to validate the tests against job performance, Bauman's work representing the sole creditable attempt.

Those psychologists, educators, and counselors who take sides on the question of the importance of tests in the guidance of the blind will find further material for discussion in this book. Hayes reports that Herbert Moore, after using a few unstandardized tests on the blind pupils of several residential schools in 1935, concluded that "measurements of tactual and motor aptitude must occupy a relatively subordinate place in student guidance." Bauman's more recent experience in a guidance clinic convinced her that "Tests can be of great value in the selection of the job in which a chosen individual shall be placed."

Hayes did not choose to include in this book an account of the efforts made to date to adapt interest and personality inventories for the blind. However, surveys on intelligence and achievement tests, and on personality inventories have been covered in other publications of the author. His most recent review was given in a paper presented at the 37th Convention of the American Association of Instructors of the Blind in 1944, under the title, "What's New In Testing the Blind?"

This newest book by an experimenter who has worked almost alone for three decades in a psychology of the blind ought to be read by all professional persons interested in the education, guidance and vocational rehabilitation of the blind. Clinical and Industrial Psychologists will learn from this book where they may find source information—first, on the experimental adaptations for the blind of some tests now widely used with the sighted, secondly, on the literature which describes past attempts to construct special tests for the blind. Research workers interested in the construction of a battery of vocational aptitude tests for the visually impaired will find excellent leads for further study. Graduate students should have little trouble in locating a wealth of "problems" that have both academic and practical significance.

The reader may be disconcerted to learn that so few have turned their energies to the vocational testing of the blind. It is hoped that the book will stimulate more systematic, widespread and coordinated endeavors to conduct studies in this field. Such investigations will make contribu-

tions not only to theoretical knowledge but also to human and social values. A psychology of the blind gives fresh insight into the psychology of all human personality.

SALVATORE G. DiMICHAEL

*Office of Vocational Rehabilitation,
Washington, D. C.*

Lowenfeld, Berthold. *Braille and talking book reading: a comparative study*. New York: American Foundation for the Blind, 1945. Pp. 53.

This monograph describes a scholarly attempt to determine the practical usefulness of a new educational device for the visually handicapped.

For blind adults the talking book is an unqualified success—a gift from heaven. Although invented less than fifteen years ago, there are already nearly 30,000 machines in use, furnishing recreational reading for many who have never mastered the slow and difficult process of reading with the fingers, while widening the intellectual horizon for those who can read braille. Educators of blind children are debating the wisdom of an extensive use of the device in school work. Is there a danger that children might refuse to learn to read braille? Will their spelling suffer? Certainly a wider acquaintance with literature, social studies and science would result, compensating for the limitations imposed by finger reading, which averages only about one-third as fast as reading with the eyes. And the superior pronunciation of the professional readers whose voices come from the records might promote good speaking habits in the listeners.

The experiments reported in this monograph were planned to compare the speed and comprehension of talking book reading with braille reading, and incidentally to throw some light upon the children's own preferences in talking book material.

Carefully controlled experiments were made in which the McCall-Crabbs Standard Test Lessons in Reading were presented to 481 pupils in grades three, four, six and seven in twelve residential schools for the blind. The material was given in braille and by the use of three different kinds of talking book records—simple readings, readings with sound effects dubbed in from sound effect records, and readings with dramatizations performed by experienced actors. The tests included stories and factual textbook material. Comprehension was measured by multiple choice questions phrased to test understanding rather than mere rote memory. After the tests had been completed the children were asked to list the four stories they liked best in order of preference.

Wide differences in rate of braille reading were found in the different schools tested, but for the whole group studied, the rate for braille reading

was only about one-third the rate at which the material was presented on the talking book. If comprehension is satisfactory a clear case would seem to have been made for the use of the talking book, since three times as much material would be covered.

At the third and fourth grade level, comprehension of straight talking book reading is significantly superior to braille reading. At the sixth and seventh grade level no significant difference in comprehension was noted for stories, while comprehension of textbook material was significantly better when presented in braille. The author suggests that this superiority of braille reading may be explained by "past practice and habituation" and that practice with the talking book and the development of purposive listening techniques might well change the relative success of the two methods. Sound effects and dramatization added greatly to the children's pleasure but did not improve their comprehension scores.

The following recommendations of the author seem fully justified by the results of the experiments:

1. "Since pupils on the third and fourth grade level read about three times as fast by the talking book as in braille and since comprehension of talking book reading is superior to that of braille reading, the use of the talking book at this level is strongly recommended in order to compensate at least in part for the slowness of braille reading."

2. "Sound effects and dramatizations used in connection with talking book reading are an attractive feature, the use of which is suggested to stimulate reading interest in blind pupils."

3. "On the sixth and seventh grade level where pupils have acquired some proficiency in braille reading, the use of the talking book is recommended because its rate, which at this level is about two and a half times as fast as that of braille reading, will permit much wider reading." But the author suggests that informational material for which the fullest possible comprehension is essential be read in braille.

4. "Many pupils of lower intelligence never achieve any real proficiency in braille reading in spite of long and laborious instruction and practice. The use of the talking book is particularly recommended for these pupils."

SAMUEL P. HAYES

*Perkins Institution and
Mass. School for the Blind*

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to
Donald G. Paterson, Editor, Department of Psychology, University
of Minnesota, Minneapolis 14, Minnesota

- Non-projective personality tests.* Harold A. Abramson, et al. New York: The New York Academy of Sciences, 1946. Pp. 678. \$1.75.
- Executive practices in the field of human resources.* Bulletin No. 12. Lawrence A. Appley. California: Industrial Relations Section, California Institute of Technology, 1946. Pp. 24. \$.50.
- Mutual survival, the goal of unions and management.* E. Wight Bakke. Connecticut: Labor and Management Center, Yale University, 1946. Pp. v + 82. \$1.00.
- Psychology applied to personnel.* Henry Beaumont. New York: Longmans, Green, and Co., Inc., 1946. Pp. 167. \$1.75.
- Insight and personality adjustment.* Therese Benedek. New York: The Ronald Press Co., 1946. Pp. 325. \$4.00.
- Job placement of the physically handicapped.* Clark D. Bridges. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 329. \$3.50.
- The physically handicapped worker in industry.* Bulletin No. 13. Gilbert Brighthouse. California: Industrial Relations Section, California Institute of Technology, 1946. Pp. 54. \$2.00.
- Labor unions and the community.* Fannia M. Cohn. New York: Workers' Education Bureau, 1946. Pp. 11. \$.10.
- Forecasting college achievement.* Albert B. Crawford and Paul S. Burnham. New Haven: Yale University, 1946. Pp. 291. \$3.75.
- How to find and succeed in your postwar job.* Frank S. Endicott. Pennsylvania: International Textbook Co., 1946. Pp. 147. \$1.75.
- Psychodrama in an evacuation hospital.* Ernest Fantel. New York: Beacon House, 1946. Pp. 23. \$1.50.
- University extension and workers' education.* Alfred P. Fernbach. Indiana: National University Extension Association, Indiana University, 1945. Pp. 32. \$.25.
- Experimental hypertension.* William Goldring, et al. New York: The New York Academy of Sciences, 1946. Pp. 179. \$3.75.
- Fundamental patterns of maladjustment—The dynamics of their origin.* Lester Eugene Hewitt and Richard L. Jenkins. Illinois: State of Illinois, 1946. Pp. 110.

- Scientific vocational guidance and its value to the choice of employment work of a local education authority.* E. P. Hunt and Percival Smith. Alabama: City of Birmingham Education Committee, 1944. Pp. 80.
- The veterans' program: A complete guide to its benefits, rights and options.* Charles Hurd. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 267. \$2.00.
- Music in medicine.* Sidney Licht. Boston: New England Conservatory of Music, 1946. Pp. 132. \$3.00.
- Lincoln's incentive system.* James F. Lincoln. New York: The McGraw-Hill Book Co., Inc., 1946. Pp. 192. \$2.00.
- Origins and development of group psychotherapy.* Joseph I. Meiers. New York: Beacon House, 1946. Pp. 44. \$1.75.
- Psychodrama.* Volume I. J. L. Moreno. New York: Beacon House, 1946. Pp. 429. \$6.00.
- Student's manual to accompany "Psychology."* Norman L. Munn. Boston: Houghton Mifflin Co., 1946. Pp. 173. \$1.00.
- Men at Work.* C. A. Oakley. London: University of London Press Ltd., 1945. Pp. 301.
- Sabbatical years with pay.* Albert Persoff. Los Angeles: The Charter Co., 1945. Pp. 144. \$2.50.
- The group method in the treatment of psychosomatic disorders.* Joseph H. Pratt. New York: Beacon House, 1946. Pp. 10. \$1.25.
- The psychology of normal people.* Joseph Tiffin, Frederic B. Knight and Eston Jackson Asher. Boston: D. C. Heath and Co., 1946. Pp. 581. \$3.25.
- High school personnel work today.* Jane Warters. New York: McGraw-Hill Book Co., Inc., 1946. Pp. 277. \$3.00.
- Changing your work?* J. Gustav White. New York: Association Press, 1946. Pp. 210. \$2.50.
- Industrial psychology and personnel practice.* Industrial Welfare Division. Australia: Department of Labour and National Service, 1946. Pp. 47.

Erratum

In the October 1946 issue of the *Journal of Applied Psychology*, an error occurred in the article, "Age of Starting to Contribute versus Total Creative Output" by Harvey C. Lehman. On page 466, line 13 read, "important contributions as late as age 20," whereas it should have read "important contributions as late as age 30."

